



---

# ***Method development in the area of multi-block analysis focused on food analysis***

PhD thesis 2016

Alessandra Biancolillo

Department of Food Science

Faculty of Science

University of Copenhagen

**Title:** Method development in the area of multi-block analysis focused on food analysis

**Submission:** September 2016

### **Supervisors**

Prof. Tormod Næs

Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway

Department of Food Science, Faculty of Science, University of Copenhagen, Denmark

Prof. Rasmus Bro

Department of Food Science, Faculty of Science, University of Copenhagen, Denmark

Dr. Ingrid Måge

Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway

### **Opponents**

Assoc. Prof. Thomas Skov

Department of Food Science, Faculty of Science, University of Copenhagen, Denmark

Prof. Qannari El Mostafa

ONIRIS, USC, Sensometrics and Chemometrics Laboratory, Nantes, France

Prof. Lars Nørgaard

Institut for Naturvidenskab og Miljø, Roskilde University, Denmark

This thesis is dedicated to whoever will read it with interest and to who will hold it with pride.

*Questa tesi e' dedicata a chiunque la leggerà con interesse ed a chi la stringerà in mano con orgoglio.*

"And once the storm is over you won't remember how you made it through, how you managed to survive. You won't even be sure, in fact, whether the storm is really over. But one thing is certain. When you come out of the storm you won't be the same person who walked in."

*"Quando la tempesta sarà finita, probabilmente non saprai neanche tu come hai fatto ad attraversarla e a uscirne vivo. Anzi, non sarai neanche sicuro se sia finita per davvero. Ma su un punto non c'è dubbio. Ed è che tu, uscito da quel vento, non sarai lo stesso che vi era entrato."*

**Haruki Murakami**

# Preface

This thesis is submitted in order to obtain the PhD degree from the PhD school of Science, University of Copenhagen.

The work is based on a project funded by the Research Council of Norway (Project number: 225096).

The work belongs to a joint project between the Norwegian research institute for food and fishery (NOFIMA) and the University of Copenhagen (KU). The main part of the research was conducted at Nofima, in Ås, Norway (approximately 70% of the work), and the rest at KU, at the food department. The entire work took three years, from September 2013 to September 2016. The PhD research was under the main supervision of Prof. Tormod Næs (Nofima, KU) and the co-supervision of Prof. Rasmus Bro (KU) and Dr. Ingrid Måge (Nofima).

The present thesis was conceived in order to develop and extend some existing statistical/chemometric tools and to propose novel methods in the multi-block field with a special focus in food analysis. The idea was to develop methods leading at the same time to good and reliable predictions and easy interpretation of data.

The present thesis is divided in two parts. In part I, six chapters are present. Chapter 1 contains an introduction to the work and its aim. In Chapter 2, a general overview of the multi-block field and of the approaches for data analysis adopted in the present research is given. Chapter 3 presents an introduction to classification and a description of the mostly used methods in this research. Chapter 4 is about variable selection, with a specific focus on methods that have been applied in the work. Chapters 5 and 6 present the conclusions and paper summaries, respectively. Part II includes four papers. The first two (Paper I and Paper II) are published, Paper III is submitted and Paper IV is a manuscript (tentative submission date: March 2016).

# Acknowledgements

First of all, I would like to acknowledge the Research Council of Norway for financing the project I have been working on in these years and made this great experience possible.

It's incredible to realize that three years have already passed! They ran so fast because I have been lucky with the people I encountered on my path. A special thanks is for the three people who supervised me during this period: Tormod Næs, Ingrid Måge and Rasmus Bro. You have been a great team! You all have always been available and helpful under any aspect of my working (and not) life in these three years. I would not only thank you all for all the scientific discussion, but in particular because you have always helped me being "in the track". Any of you, in your own way, helped me to take the right direction without ever feeling lost; Thanks!

I would express my gratitude to all the people in the Råvare og prosess group and in particular to Dr. Ragni Ofstad, because they kindly welcomed me and they let me felt, from the very beginning, part of their big family. I consider a privilege to have worked in such a nice environment, and everyone is contributing in making it so special.

A big thanks goes to some people I have collaborated with, like Kristian Liland and Kasper Christiansen, in particular for the exchange of nerd stuff and knowledge ☺

I would thank all the PhD-students I have met in these years, for sharing hints, tricks and fears; in particular the ones in NOFIMA, I will always have special memories of the PhD trip we had together throughout Norway.

My gratitude goes to a special friend and room-mate, Dimitrios, for all the chats and the support; in particular for being the man of the wise-advises about life and gardening ☺ A big thanks goes also to a lovely couple of friends, Marta and Joaquin, and to Georgely, Kasper, Jens, Karen and Daniele for the nice time spent together. Winter can be tough up there, but you have been a safe shelter in the darkest moments (Sorry to tell you, but I have to admit that also hot-chocolate helped a lot! ☺ ).

I would thank all the people at KU in SPECC aka CAT aka whatever-it-will-be at the time you will be reading this thesis considering the frequency which the group is changing name ☺ Unfortunately, I haven't spent so much time in Copenhagen, and sometimes my visits have

been very short. Despite this, I have never had the feeling I was a stranger in the group; I really really appreciated it.

Finally, I would thank my family and my Italian friends; no matter how strong the storm has been, they have always been present, even from far.

# Table of contents


Preface.....	i
Acknowledgements.....	ii
Table of contents.....	iv
Abstract.....	1
Resumé.....	2
Riassunto.....	3
List of Papers.....	4
List of figures.....	4
Abbreviations and acronyms.....	5
Notation and List of Mathematical entities.....	6
<b><u>Part I:</u></b> .....	<b>8</b>
<b>Chapter 1    Introduction.....</b>	<b>9</b>
1.1 Food Quality and Chemometrics.....	9
1.2 Aim of the present work.....	11
<b>Chapter 2    Multi-block Regression.....</b>	<b>14</b>
2.1. The background of multi-block regressions.....	14
2.2 Data structures useful for multi-block models.....	17
2.3 Main multi-block methods used in this work.....	18
2.3.1 Brief description of Multi-block-PLS (MB-PLS).....	18
2.3.2 Brief description of Parallel Orthogonalized-PLS (PO-PLS).....	19
2.3.3 SO-PLS in detail.....	20
2.3.3.1 Estimating the optimal number of latent variables in SO-PLS.....	24
2.3.3.2 Interpretation of SO-PLS models.....	30
2.3.4 SO-N-PLS in detail.....	32
2.3.4.1 Interpretation of SO-N-PLS models.....	35
<b>Chapter 3    Multi-block Classification.....</b>	<b>36</b>

3.1 Brief introduction to classification.....	36
3.2 Linear discriminant analysis (LDA).....	37
3.3. PLS-based classification methods.....	40
3.3.1 SO-PLS-LDA in detail.....	40
3.3.1.1 Definition of the optimal number of latent variables in SO-PLS-LDA.....	40
3.3.2 SO-N-PLS-LDA in detail.....	41
3.3.3 Interpretation of classification results in the SO-PLS-based models.....	41
<b>Chapter 4 Variable Selection.....</b>	<b>44</b>
4.1 Variable selection methods used in the work.....	46
4.2 Introduction of a variable selection step in multi-block models.....	51
<b>Chapter 5 Conclusions.....</b>	<b>55</b>
5.1 Main results.....	55
5.2 Future Perspectives.....	57
<b>Chapter 6 Paper summaries.....</b>	<b>59</b>
6.1. Summary of Paper I.....	59
6.2. Summary of Paper II.....	59
6.3. Summary of Paper III.....	59
6.4. Summary of Paper IV.....	60
6.5 References.....	61
<b><u>Part II:</u>.....</b>	<b>67</b>
<b>Publications.....</b>	<b>68</b>



## Abstract

In data analysis one could be interested in the relations among a number of data sets (data blocks) having different origin. In food science, this can be particularly relevant. For instance, developing a new product, one may need to understand the relation between physical/chemical data, sensory data and consumer acceptance data. A further example could be in process monitoring, where one of the main tasks is to figure out relations among spectroscopic measurements on raw materials and/or during the production, process settings, and the quality of end-product(s). Additionally, data blocks could have not only different origin, but measurements could be taken at different time points or by multi-channel instruments. It has been demonstrated, that it is more convenient to extract information from multi-block data sets handling all the blocks at the same time. Namely, performing *data fusion* by the means of multi-block methods. Several statistical and chemometric multi-block methods are already available. Mainly, these are natural developments and variations of previously widely-used methods in multivariate analysis, but the area still needs to be explored. This PhD project is centered on method-development and method-testing in the multi-block analysis field, with a specific focus on food analysis. Novel approaches will be compared with other well-known methods used in the same field and they will be applied both in regression and in classification. The new methodologies will be tested on simulations and on real data. Attention will be also given to categorical input data (*Paper IV*). Additionally, variable selection in this context will be investigated, in order to obtain reduced sub-sets, easier to interpret (*Paper II*). In conclusion, due to the increasing need of handling multi-way arrays ( i.e., structures resulting from experiments where the data are collected as a function of more than two sources of variability), all the considerations done in the first part of the study, will be extended to multi-way arrays (*Paper III*).



***This PhD project is centered on method-development and method-testing in the multi-block field, with a specific focus on food analysis.***

# Resumé

I dataanalyse kan man være interesseret i at undersøge sammenhængen mellem en række datasæt (datablokke) af forskellig art. I fødevarevidenskab, kan dette være særlig relevant. For eksempel, kan man ønske at forstå sammenhængen mellem fysisk/kemiske data, sensoriske data og forbrugernes accept af data, når man udvikler et nyt produkt. Et andet eksempel kunne være i procesovervågning, hvor en af de vigtigste opgaver er at finde relationerne mellem spektroskopiske målinger på råvarer, procesindstillinger, og kvaliteten af slutproduktet. Blokke af data kommer ikke kun fra forskellige målinger. Det kan også komme fra at man måler på forskellige tidspunkter eller ved flere kanaler i et instrument. Det er blevet vist, at det kan give mere information fra når man ser samlet på sådanne blokke og dette kaldes multi-blok modellering eller data fusion.

Der findes allerede flere statistiske og kemometriske multi-blok metoder. De er ofte naturlige udviklinger og variationer over multivariate analysemetoder, men metoderne er langt mindre modent end ønskværdigt. Dette ph.d.-projekt er centreret om metode-udvikling og metode-test i multi-blok analyse, med særligt fokus på fødevareanalyse. Nye metoder vil blive sammenlignet med kendte metoder, og de vil blive anvendt både til regression og klassifikation. De nye metoder vil blive testet på simulerede såvel som på reelle data. Der vil blive set på situationer hvor man har kategoriske input data (Papir I og papir IV). Derudover vil variabel selektion undersøges, for at opnå mindre datasæt, lettere fortolkning og bedre prædiktioner (Paper II).

På grund af det stigende behov for at håndtere multi-vejs data (dvs. strukturer som følge af eksperimenter, hvor oplysningerne indsamles som en funktion af mere end to modi), vil cden udviklede metodik også blive udvidet til multi-vejs data (Paper III).

## Riassunto

Nella scienza alimentare si è spesso interessato a capire la relazione tra dati provenienti da diversi piattaforme e/o differenti ambiti. Ad esempio, nella produzione (e nello sviluppo) di un prodotto è necessario combinare dati chimici, sensoriali ed opinioni dei consumatori. Risulta quindi necessario mettere insieme informazioni provenienti da settori scientifici differenti. Nell'ottimizzazione dei processi invece, combinare dati da piattaforme differenti o da medesime piattaforme in tempi differenti, ha un ruolo fondamentale. Infatti, un controllo in continuo sul ciclo produttivo (effettuando misure chimiche sulle materie prime, nei punti critici del processo e sul prodotto finito) garantisce non solo l'efficienza della filiera ma anche la qualità del prodotto finale. Inoltre, nell'ambito delle scienze alimentari, ci si può trovare ad analizzare gli effetti dei prodotti su uomini o animali. Indagini che generalmente avvengono combinando analisi microbiologiche, chimiche (su diversi supporti, come sangue, feci ed urina), dati di espressione genica e descrittori di salute. Queste misurazioni sono spesso presi in diversi punti temporali, e generalmente mediante tecniche completamente diverse e relative a settori scientifici differenti. In tutte queste situazioni, diversi blocchi di informazioni necessitano di essere elaborati insieme al fine di estrarre le informazioni. Si tratta infatti dei cosiddetti "Multi-block" data sets. Oggigiorno, sono numerosi i metodi chemiometrici/statistici disponibili per combinare i dati provenienti da diverse piattaforme. Questi metodi sono i così detti metodi di *data fusion* (o metodi *Multi-block*). La maggior parte di queste metodologie sono naturali estensioni di metodi di analisi multivariata (chemiometrici e statistici) precedentemente consolidate; ma il campo è ancora ampiamente sotto studio. Molte sono infatti ancora le lacune che affliggono questo ramo dell'analisi dati. Il presente progetto si situa in questo campo, al fine di sviluppare nuove metodologie analitiche multi-block finalizzate in particolare all'analisi nell'ambito alimentare. In particolare, in questo studio sono stati realizzati un metodo multi-block e multi-way per la regressione (SO-N-PLS e due per la regressione (SO-PLS-LDA/ SO-N-PLS-LDA). Inoltre, la rappresentazione grafica dei risultati è stata studiata ed è stato proposto un mezzo investigativo per i risultati in classificazione. Nella parte finale del lavoro, l'interpretazione di alcuni metodi multi-block è stata discussa, ma le considerazioni fatte hanno bisogno di un maggiore studio (e di validazione su dati reali) prima di poter essere considerate attendibili.

## List of Papers

- I. A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, *Chemometrics and Intelligent Laboratory Systems*, 141 (2015) 58–67.
- II. A. Biancolillo, K. Hovde Liland, I. Måge, T. Næs, R. Bro, Variable selection in multi-block regression, *Chemometrics and Intelligent Laboratory Systems*, *Chemometrics and Intelligent Laboratory Systems* 156 (2016) 89–101.
- III. A. Biancolillo, T. Næs, R. Bro, I. Måge Extension of SO-PLS to multi-way arrays: SO-N-PLS. *Submitted to Chemometrics and Intelligent Laboratory Systems on 29<sup>th</sup> June 2016*
- IV. A. Biancolillo, T. Næs, R. Bro, I. Måge, Multi-block regression: combining data sets with different structures and dimensionalities-*Manuscript*

## List of Figures

- Figure 1: Conceptual flow chart: Overview
- Figure 2: Conceptual flow chart: Multi-Block Regression
- Figure 3: Origin of multi-block data sets
- Figure 4: Summary of SO-PLS's algorithm
- Figure 5: Latent variables selected in the Simulations Study
- Figure 6: Måge Plot for SO-PLS
- Figure 7: Conceptual Flow chart: Multi-block Classification
- Figure 8: Proposed graphical representation tools for SO-PLS-LDA models
- Figure 9: Conceptual Flow Chart: Variable selection

# Abbreviation and Acronyms

Abbreviations and acronyms frequently used in the text

**ANOVA:** ANalysis Of VAriance

**CV:** Cross-Validation

**DQ<sup>2</sup>:** Discriminant Q<sup>2</sup>

**LDA:** Linear Discriminant Analysis

**LV:** Latent Variables

**MB-PLS:** Multi-Block Partial Least Squares

**N-PLS:** N-Partial Least Squares

**PLS:** Partial Least Squares

**PO-PLS:** Parallel and Orthogonalized Partial Least Squares

**RMSECV:** Root Mean Squared Error in Cross Validation

**RMSEP:** Root Mean Squared Error in Prediction

**RSS:** Residual Sum of Squares

**sMC:** Significance multivariate correlation

**SO-N-PLS:** Sequential and Orthogonalized N-Partial Least Squares

**SO-N-PLS-LDA:** Sequential and Orthogonalized N-Partial Least Squares- Linear Discriminant Analysis

**SO-PLS:** Sequential and Orthogonalized Partial Least Squares

**SO-PLS-LDA:** Sequential and Orthogonalized Partial Least Squares- Linear Discriminant Analysis

**SR:** Selectivity Ratio

**TSS:** Total Sum of Squares

**UVE:** Elimination of Uninformative Variables for multivariate calibration

**VIP:** Variable Importance in Projection

# Notation and list of mathematical entities

## Notation

In the text, matrices are indicated by bold capitals (e.g. **X**), multi-way arrays by underlined bold capitals (e.g. **X**) and vectors by bold lowercase letters (e.g. **x**). Scalars are indicated by italics (e.g., *N*).

The symbols  $\mathbf{X}_{(N \times J)}$ ,  $\mathbf{Z}_{(N \times H)}$ ,  $\underline{\mathbf{X}}_{(N \times J \times K)}$ ,  $\underline{\mathbf{Z}}_{(N \times H \times L)}$ ,  $\mathbf{X}_{\text{un}(N \times JK)}$  and  $\mathbf{Z}_{\text{un}(N \times HL)}$  represent predictor matrices, three-way array predictors, and unfolded three-way predictors, respectively. The response matrix is represented by  $\mathbf{Y}_{(N \times A)}$ , while the response vector is  $\mathbf{y}_{(N \times 1)}$ .

In Part I, the following letters refer always to the same entity:

*A*: Number of columns in the response matrix

**B**, **b**: Regression coefficient matrix, Regression coefficient vector

*C*: canonical variate scores

**E**: Error matrix

*F*: Number of components

*G*, *g*: Total number of groups, Individual group (in classification)

*H*: Number of variables in the **Z** block; number of variables in the second mode of the **Z** array

*J*: Number of variables in the **X** block; number of variables in the second mode of the **X** array

*K*: Number of variables in the third mode of the **X** array

*L*: Number of variables in the third mode of the **Z** array

*N*, *n*: Total number of samples, sub-set of samples

**P**: **X**-Loadings in PLS regressions

**Q**: **Y**-Loadings in PLS regressions

**R**: weights matrix in SO-N-PLS

**S**: Variance/covariance matrix

**T**: **X**-Scores in PLS regression

**U:** Y-Scores in PLS regression

**V:** weights in PLS regression (relating directly scores **T** to un-deflated **X**)

**W:** weights in PLS regression

### Mathematical entities

**Classification Error:** Percentage of samples misclassified

$$E_{class} = \frac{n_{miscl}}{N} \cdot 100$$

**R<sup>2</sup>/ Q<sup>2</sup>:** Coefficient of determination (Calibration/Prediction)

$$R^2(Q^2) = 1 - \frac{RSS}{TSS}$$

**RMSECV:** Root Mean Square Error in Cross-Validation

$$RMSECV = \sqrt{\frac{\sum_{n=1}^N (y_n - \hat{y}_{n,CV})^2}{N}}$$

**RMSEP:** Root Mean Square Error in Prediction

$$RMSEP = \sqrt{\frac{\sum_{t=1}^{N_t} (y_t - \hat{y}_t)^2}{N_t}} \quad \text{with } N_t \text{ number of samples in the test set}$$

**RSS:** Residual Sum of Squares

$$RSS = \sum_{j=1}^N (y_j - \hat{y}_j)^2$$

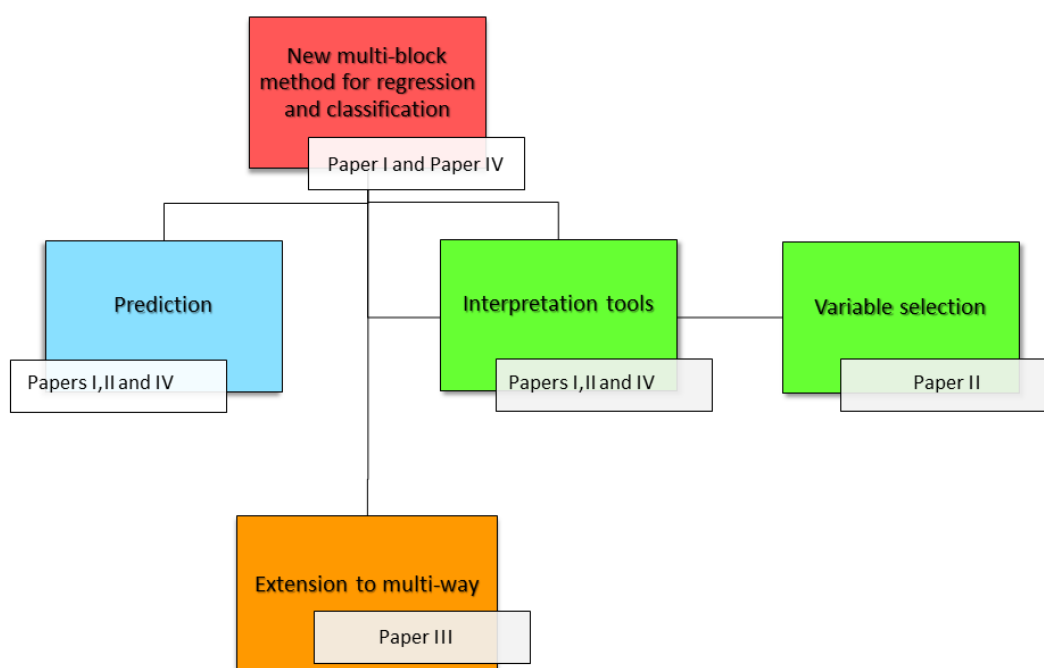
**TSS:** Total Sum of Squares

$$TSS = \sum_{j=1}^N (y_j - \bar{y})^2$$

# Part I



# Chapter 1: Introduction



**Figure 1: Conceptual flow chart: Overview among topics discussed in the thesis.**

## 1.1 Food Quality and Chemometrics

Food quality is a highly relevant topic in modern society; it embraces several important aspects and is a field of study that involves a large number of different disciplines. A general awareness of the importance of having foodstuffs that reflect certain quality standards, is therefore widely spread. These standards do not involve only sanitary features, but also nutritional and hedonistic aspects. The first definitions of food quality are actually quite old. One pioneer of quality management, M. Juran defined “*quality*” as “*those features of products which meet customer needs and thereby provide customer satisfaction*” [1]. Nowadays, standards of quality for foodstuff are strictly regulated by national or international institutions. For instance, EU has enacted several laws to define quality and to protect consumers (like ISO 9000:2005 and EC 628/2008 just to mention a couple of them).

The aim of regulating foodstuff production is not only to have common (national and international) rules but it is also to preserve characteristics of traditional food. As a consequence, authentication and traceability of foodstuff have become hot topics. In this context, chemometrics has an extremely relevant role. In fact, in combination with analytical techniques, it represents one of the most powerful ways for the investigation of foodstuff.

Among all the physical-chemical techniques used in this field, chromatography and spectroscopy are the most common. In particular, the latter is relatively cheap, fast and non-destructive. Consequently, spectroscopic techniques are well suited for online measurement (to inspect intermediate products in the chain) and to check the quality of end-products. Several approaches have been developed to combine these techniques with chemometrics in order to predict key quality attributes or classify different product qualities. Chemometric tools can be used to authenticate products, check for frauds or adulteration and also monitor the productive processes, avoiding deficiencies in the final product [2-3]. In particular, classification models are widely used in the traceability of the food chain assessing the geographical origin of products (or of sub-products). Some examples can be found in [4-7]; these represent only a small part of the wide number of papers published on this regard. In this context, data fusion or multi-block approaches play an important role. Their power belongs to their ability of exploiting the underlying relation between all or several of the data blocks involved. Often data come naturally as *multi-block data sets* (also called *multi-sets*). For instance, in process monitoring different measures can be taken on several batches over time, generating a multi-block and multi-way set of data. Multi-block data sets are also generated when different measures are collected on the same production lot or when a single technique is applied on several batches. Additionally, a recipe (presence/absence or concentrations of ingredients), assessments on the final product quality and/or consumers acceptance could be collected together in order to understand the relation among them; also this one represent a multi-set of data. In all these cases, it is more effective to extract info handling all the blocks at the same time rather than building individual models for each set of data [8-9].

Several multi-block methods have been developed for regression and classification purposes and they are widely applied in food analysis. Data fusion can be performed at different *levels*: *measurements-level* (also known as *low-level data fusion*) and *components-level* (*mid-level data fusion*). In the first case, the extraction of features is performed after the concatenation of data (e.g., as in Multi-block PLS). Instead, in the *component-level* data fusion, features are extracted individually from the blocks and then used for a joint analysis in successive stages. Despite the level of the data fusion, multi-block methods show generally good performances, with high accuracy in prediction. Anyhow, some aspects still deserve research.

As already stated, chromatography and spectroscopy are analytical techniques widely used in food analysis. Often their outcomes have several regions (of the spectra/chromatogram) that are not relevant for the current prediction or classification problem. These regions could

negatively affect (both regression or classification) models. This negative influence could even be enhanced in multi-block data sets (e.g. having chromatograms and spectra as predictor blocks). In order to avoid this effect, a valid solution would be to apply a *variable selection method* to reduce the number of variables. Getting rid of noise and redundancy by variable selection, predictions may improve and the model become easier to interpret. Consequently, the inclusion of a variable selection step building (regression or classification) models, would definitely represent a valid contribution to food analysis. Variable selection is not a straightforward task; it is difficult to define which tool would be the most suitable with a specific data set. How to proceed in a multi-block data context is even more complex. Consequently, how to combine variable selection tools with multi-block methods is definitely a topic worth to research.

Another topic that is becoming more and more interesting is the development of multi-block methods for multi-way data arrays. As mentioned, in food analysis it is common to encounter data sets constituted by multi-way arrays (i.e., structures resulting from experiments where the data are collected as a function of more than two sources of variability). For instance, in sensory science assessments are reported as functions of the sample, the judge and the attribute. In process monitoring variables may be measured on different batches along time. Also, new technologies have made available analytical instruments that can collect multi-way arrays of data (e.g., fluorescence or Nuclear Magnetic Resonance). Different method has been developed to handle this type of data, but only few can handle several predictor arrays [10-13]. So far, the common practice is to *unfold* (i.e. make the array two-way) the multi-way arrays, but this presents some issues (more detail in Paragraph 2.2). Therefore, multi-block methods for multi-way arrays is definitely a need in food analysis.

## **1.2 Aim of the present work**

The aim of the present work is to develop and test new regression methods in the multi-block field, with a special focus on food analysis; both prediction ability and interpretation will be of interest.

Although, as reported in the previous paragraph, several multi-block algorithms have already been proposed in the literature, still there are many issues that need to be examined in greater depth. In this context, the attention will be focused on a particular kind of multi-block approach, namely Sequential and Orthogonalized PLS (SO-PLS) [14]. This method constitutes the basis of almost all research carried out in the present thesis. As it will be more evident in Chapter 2, where the algorithms are described in depth, SO-PLS presents a number

of advantages when compared to other existing methodologies. These benefits represent the reason why this method has been chosen as the starting point for the development of the novel methods. One of these advantages is its sequential nature, which allows evaluating whether the additional contribution of successive blocks is relevant and, if so, to interpret it. This characteristic represents a peculiarity, not only if compared to methods which rely only on the extraction of components which are common among the blocks (such as Multi-Block-PLS (MB-PLS) [15] or Common Components and Specific Weight Analysis (CCSWA) [16]), but also with respect to other approaches which use differently the block-specific (distinctive) information (e.g., OnPLS [17] or Parallel Orthogonalized-PLS (PO-PLS) [18]). Moreover, this makes SO-PLS a suitable method for both prediction and interpretation purposes.

Applications will focus on predictions of chemical constituents and sensory measurements from different analytical techniques.

In food data analysis it may also be highly relevant to classify objects, in order to identify, interpret and visualize classes of individuals. This can be useful in process monitoring, in order to check, at different steps of the production process, if there are deviations from the usual characteristics of the product. Therefore, multi-block classification will also be focused in this work by the same regression methodologies as used in the rest of the thesis. In particular, attention will be given to interpretation and visualization tools. In order to avoid overfitting and overoptimistic interpretation, these will be based on cross-validation (Paper I). Moreover, starting from the assumption that a reduced model is more easily interpretable, another goal of the present work is to investigate variable selection in the context of multi-block analysis. This is important not only for interpretation but also for selecting simplified sub-sets of variables to be used in future studies. This issue is particularly demanding when the number of variables is larger than the number of objects, and becomes even more challenging in a multi-block framework, as more matrices are involved and their mutual relation should be taken into account (Paper II).

All the considerations done in the first part of the work, will be extended and discussed for multi-way data. As a consequence, the possibility of opening the multi-block field to multiway arrays is investigated and a new method (applicable both in regression and in classification) is proposed (Paper III). In the final paper (paper IV) the aim is to investigate in more detail interpretability of (some) multi-block methods.



***Main emphasis in the method-development is on:***

- Classification
- Graphical representation
- Variable selection
- Multi-way input blocks
- Interpretation

## Chapter 2: Multi-block regression

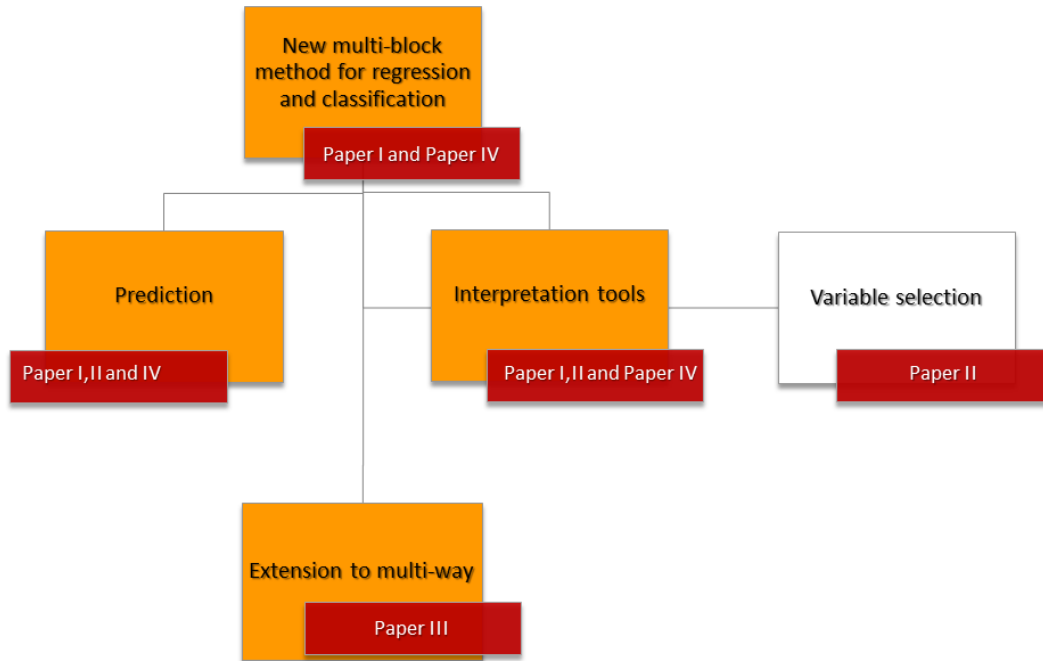


Figure 2: Conceptual flow chart: Multi-block Regression

### 2.1 The background of Multi-block regression

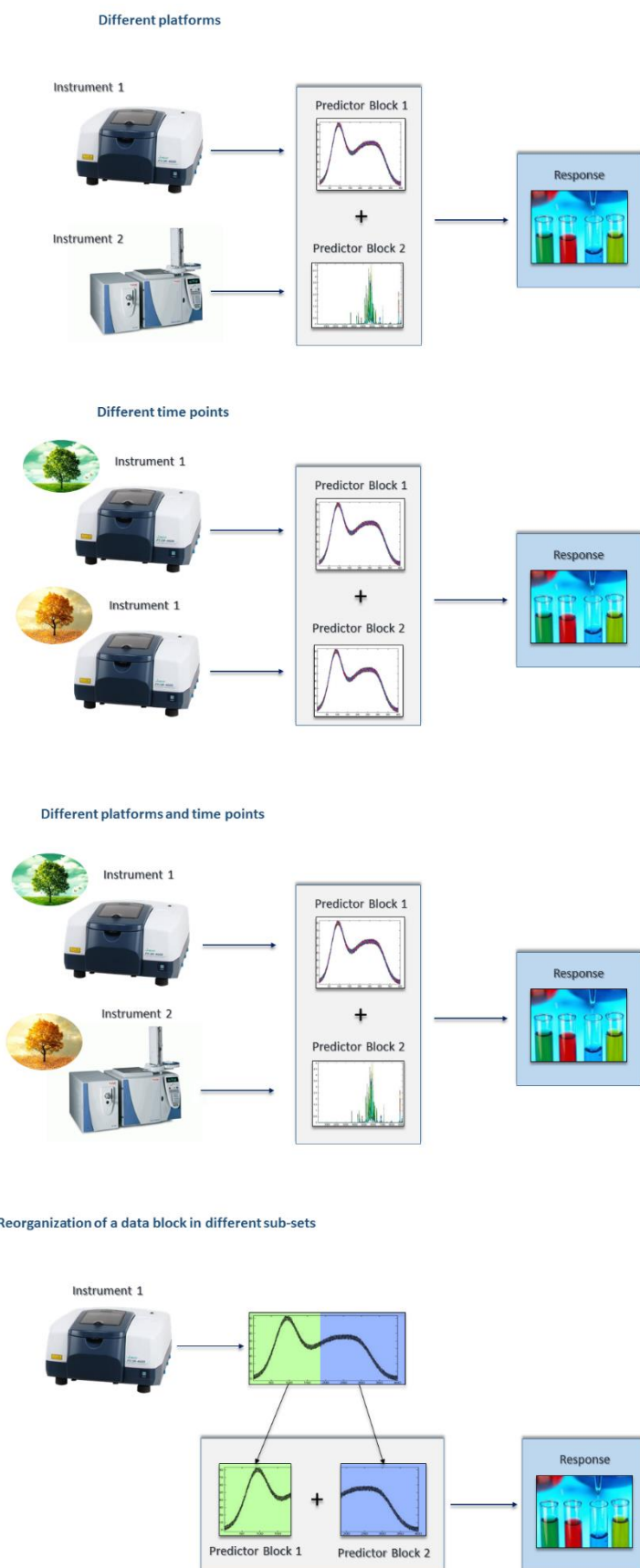
Multi-block methodologies can be split into two different parts, regression oriented approaches with a predictive direction among the blocks and approaches with no predictive direction. The first are named *multi-block regression problems* and the other *multi-block component problems*. A multi-block regression problem is present when different data matrices are used to predict one or more responses. In multi-block component problems there are different blocks of data (with at least one common mode) and the components which summarize information in them are extracted and investigated. The extraction of information can be done in various ways; different methods (e.g., CPCA, DISCO, JIVE, CCSWA and all the variants of ComDim [16,19-23]) have been proposed.

As indicated above, multi-block data sets can be obtained for different reasons (summarized in Figure 3) and in different fields. Some data sets are “intrinsically” multi-block, as the experiments *per se* generate a set of different matrices. For instance, data can be collected from different platforms or at different time points (or both); resulting in diverse sets of measures on the same samples.

Additionally, data sets can become multi-block because of a reorganization of their structures. For example, an individual block of data could be rearranged in multiple sub-sets to avoid the

number of variables exceeding the number of objects, or it could be reorganized in sub-groups because of a priori knowledge (e.g., a spectroscopic measurement taken online, could be split in different blocks according with time).

Currently, there are many methods available to solve multi-block regression problems, such as Hierarchical-PLS [24], Multi-Block-PLS (MB-PLS) [15,25-29], Sequential and Orthogonalized Partial Least Squares (SO-PLS) [14,30], Parallel Orthogonalized Partial Least Squares (PO-PLS)[18,31], On-PLS [17], Multiblock Redundancy Analysis [32-33], Predictive-ComDim [23]; and these are just examples of some of them. In spite of this, multi-block is a quite new area, and many aspects still need to be inspected.



**Figure 3** *Origin of multi-block data sets*



## 2.2 Data structures useful for multi-block models

### *Two, three and multi-way arrays*

The main part of the multi-block methods handle one- or two-way arrays, but in practice it is quite common to have data organized in three-(or more) way arrays. As mentioned in Paragraph 1.1, examples of this can be easily found in process monitoring or in sensory science. Moreover, modern instruments often produce multi-way arrays. Finally, different type of data can be re-organized in higher order arrays. Consequently, the lack of multi-block methods handling these data structures is a huge deficiency.

Despite some methods have been proposed to handle higher order arrays [10-13] they are still a minority. The most common procedure dealing with multi-way arrays is to *unfold* them to structures suitable for classical multi-block methods (namely, two-way structures). The main *unfolding* procedures are three: *row-wise unfolding*, *column-wise unfolding* and *tube-wise unfolding*. Some others can be found in literature, but they can be seen as variations of these three. Starting from a three-way array  $\underline{\mathbf{X}}$  of dimensions  $N \times J \times K$ , the application of any type of unfolding reorganizes data in a two-mode arrays. The *row-wise* unfolding provides a matrix  $\mathbf{X}$  of dimensions  $N \times (JK)$ , the *column-wise* approach unfolds to a matrix  $\mathbf{X}$  of dimensions  $(NK) \times J$  and the *tube-wise* unfolding gives an  $\mathbf{X}$  matrix of dimensions  $(NJ) \times K$ .

Handling multi-block and multi-way data sets, the common practice is to perform unfolding maintaining unchanged the common mode among blocks (and re-organize the other two). Any mode of the arrays can be the common one. Often, it is the first one (samples mode); but it happens that data sets have variables in common instead of samples (i.e. psychometrics). In all the examples presented in this work, the common mode is the sample mode; consequently, when the unfolding procedure is mentioned, it is the row-wise unfolding. Unfolding procedures presents some issues that could affect the models. Therefore, development of a multi-block method which can handle multi-way sets of data appears a reasonable goal that needs to be reached.

## **2. 3 Main Multi-block methods used in this work**

As mentioned, the present work is focused on the application of the Sequential and Orthogonalized-Partial Least Squares (SO-PLS) method in the food analysis. The reason for this choice is based on a number of interesting properties of the method both with respect to interpretation and flexibility to be discussed more in detail below. At the same time, prediction performances are generally good. Moreover, it can easily handle combinations of design variables and multi-collinear predictor variables. Finally, it is computationally fast and easy to implement.

In order to fully characterize the method and its performances, SO-PLS and its extensions to be presented below will be compared to Multi-block-PLS (MB-PLS). MB-PLS has been chosen as main counterpart because it is a well-established and a well-performing method. On the other hand, in a data fusion perspective, MB-PLS and SO-PLS differ in the way the information from the various block of predictors is joined: indeed, in the framework of the distinction reported in Paragraph 1.1, MB-PLS represents an example of low-level fusion strategy, while in SO-PLS the scores are concatenated, so it can be considered a mid-level approach. Additionally, in Paper IV SO-PLS and MB-PLS are also compared with Parallel and Orthogonalized-PLS (PO-PLS). PO-PLS shares with SO-PLS some relevant features which makes the two approaches rather similar. This similarity is the reason why, in Papers I-III, PO-PLS was not selected as the method SO-PLS should be compared with. On the other hand, since the main differences between PO-PLS and SO-PLS pertain to the interpretation aspect, in Paper IV (a discussion-oriented paper focused on interpretation) it was natural to include PO-PLS in the comparison.

In the following subparagraph, MB-PLS and PO-PLS are briefly presented, while SO-PLS is discussed in detail. Here it should be noted that, in the present chapter, methods' descriptions always refer to the case of two blocks of predictors ( $\mathbf{X}$  and  $\mathbf{Z}$ ), used for the regression of the response  $\mathbf{Y}$ .

### **2.3.1 Brief description of Multi-Block-PLS (MB-PLS)**

Predictive Multi-Block PLS was proposed by Frank et alia [25-26] in 1984. Several variations and developments followed [15, 27-29].

The MB-PLS algorithm used in this work is the one presented in [29], which is slightly different from the original one [27] but provides the same model parameters.

Briefly, for two blocks  $\mathbf{X}$  and  $\mathbf{Z}$  predicting the response  $\mathbf{Y}$ , the procedure followed to build MB-PLS models is the following:

1.  $\mathbf{X}$  and  $\mathbf{Z}$  are pretreated and divided by their Froebenius' norm (obtaining  $\mathbf{X}_n$  and  $\mathbf{Z}_n$ ).
2.  $\mathbf{X}_n$  and  $\mathbf{Z}_n$  are concatenated, resulting in the matrix  $\mathbf{X}_{Conc}$  ( $\mathbf{X}_{Conc} = [\mathbf{X}_n \mathbf{Z}_n]$ ).
3.  $\mathbf{Y}$  is then fitted to  $\mathbf{X}_{Conc}$  by PLS.

As demonstrated in [15], applying PLS on  $\mathbf{X}_{Conc}$ ,  $\mathbf{X}$ - and  $\mathbf{Y}$ -scores ( $\mathbf{T}_{XConc}$  and  $\mathbf{T}_Y$ , respectively) correspond to super-scores and  $\mathbf{Y}$ -scores that would be obtained applying the original MB-PLS algorithm [27] to the same predictor blocks.

Applying Westerhuis' algorithm [15], parameters such as block weights  $\mathbf{W}_{Xb}$  and  $\mathbf{W}_{Zb}$ , super-weights  $\mathbf{W}_T$  and block scores  $\mathbf{T}_{Xb}$  and  $\mathbf{T}_{Zb}$  (present in [27]) are missing. Nevertheless, they can be calculated a posteriori. For the  $f$ -th component, it can be demonstrated that:

$$\mathbf{w}_{f,Xb} = \mathbf{X}_n^T \mathbf{t}_{f,Y} / \mathbf{t}_{f,Y}^T \mathbf{t}_{f,Y} \quad (1)$$

$$\mathbf{w}_{f,Zb} = \mathbf{Z}_n^T \mathbf{t}_{f,Y} / \mathbf{t}_{f,Y}^T \mathbf{t}_{f,Y} \quad (2)$$

$$\mathbf{t}_{f,Xb} = \mathbf{X}_n \mathbf{w}_{f,Xb} \quad (3)$$

$$\mathbf{t}_{f,Zb} = \mathbf{Z}_n \mathbf{w}_{f,Zb} \quad (4)$$

$$\mathbf{\Theta}_f = [\mathbf{t}_{f,Xb} \mathbf{t}_{f,Zb}] \quad (5)$$

$$\mathbf{w}_{f,T} = \mathbf{\Theta}_f^T \mathbf{t}_{f,Y} / \mathbf{t}_{f,Y}^T \mathbf{t}_{f,Y} \quad (6)$$

The super-weights  $\mathbf{W}_T$  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 allows the characterization of the individual blocks.

### 2.3.2 Brief description of Parallel Orthogonalized-PLS (PO-PLS)

Parallel Orthogonalized-PLS [31, 18] is a multi-block regression method which allows the identification (and the extraction) of *common* and *distinct* components among the predictor blocks. The extraction is performed in parallel among the predictors; which means this method is suitable when one is interested in common/unique information among blocks rather than in the variability added from each matrix to the model. The algorithm applied in the

present work is not the original one [31] but its variation presented in [18]. It can be summarized by the following steps:

- 1)  $\mathbf{Y}$  is predicted from  $\mathbf{X}$  and  $\mathbf{Z}$  by two individual PLS models. Scores from these models are called  $\mathbf{T}_X$  and  $\mathbf{T}_Z$ .
- 2) Common components are identified by canonical correlation analysis [34] between  $\mathbf{T}_X$  and  $\mathbf{T}_Z$ . The number of common components is decided by evaluating the values of canonical correlations. The common scores  $\mathbf{T}_C$  are the average canonical scores (from each block).
- 3) Scores  $\mathbf{T}_X$  from 1) are orthogonalized with respect to  $\mathbf{T}_C$ , giving  $\mathbf{T}_{Xorth}$ .
- 4)  $\mathbf{Y}$  is then predicted from  $\mathbf{T}_{Xorth}$  by PLS regression giving  $\mathbf{T}_{DXorth}$  (these scores represent the distinct information in  $\mathbf{X}$ ).
- 5) Scores  $\mathbf{T}_Z$  from step 1 are orthogonalized with respect to  $\mathbf{T}_C$  and  $\mathbf{T}_{DXorth}$ , giving  $\mathbf{T}_{Zorth}$ .
- 6)  $\mathbf{Y}$  is predicted from  $\mathbf{T}_{Zorth}$  by PLS regression giving  $\mathbf{T}_{DZorth}$  (these scores represent the distinct information in  $\mathbf{Z}$ ).
- 7) The final predictive model is obtained by :

$$\mathbf{Y} = \mathbf{T}_{PO}\mathbf{B}_{PO} \quad (7)$$

Where  $\mathbf{T}_{PO}$  is the concatenated matrix of the scores (  $\mathbf{T}_{PO} = [\mathbf{T}_C \ \mathbf{T}_{DXorth} \ \mathbf{T}_{DZorth}]$  ) and  $\mathbf{B}_{PO}$  is the regression coefficient matrix.

Information is extracted from both blocks at the same time (step 1). If there are no common components among the blocks, the PO-PLS algorithm is equivalent to SO-PLS (See below).

### 2.3.3 SO-PLS in detail

In Sequential and Orthogonalised Partial Least Square Regression (SO-PLS), the general multi-block linear regression problem can be represented by the equation:

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{c} + \mathbf{E} \quad (8)$$

Where:

$\mathbf{b}$  and  $\mathbf{c}$  are the regression coefficients of dimensions  $(J \times A)$  and  $(H \times A)$ , respectively.

$\mathbf{E}$  is the residual matrix of dimensions  $(N \times A)$ .

The SO-PLS algorithm is quite simple, and it is mainly divided in four steps:

- 1)  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  by PLS regression.
- 2)  $\mathbf{Z}$  is orthogonalised with respect to the scores ( $\mathbf{T}_X$ ) of the previous PLS, obtaining  $\mathbf{Z}_{Orth}$ .
- 3)  $\mathbf{Y}$ -Residuals from step 1) are fitted to  $\mathbf{Z}_{Orth}$  by PLS regression.
- 4) The full predictive model can be computed by combining the predictions of the two individual PLS models in 1) and 3).

A visual summary of the different step is presented in Figure 4.

Taking an accurate look into the different steps it is possible to find out the characteristics and the peculiarities of the method.

### **Step 1: First regression**

As said,  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  by PLS regression; this is the starting point of the method. In this step the  $\mathbf{X}$ -scores  $\mathbf{T}_X$ , the  $\mathbf{X}$ -weights and loadings  $\mathbf{W}_X$  and  $\mathbf{P}_X$  and the  $\mathbf{Y}$ -loadings  $\mathbf{Q}_X$  are obtained. The matrix of  $\mathbf{Y}$ -residuals  $\mathbf{E}$  is then calculated as:

$$\mathbf{E} = \mathbf{Y} - \mathbf{T}_X \mathbf{Q}_X^T \quad (9)$$

### **Step 2: Orthogonalization**

Given the sub-space identified by  $\mathbf{T}_X$ ,  $\mathbf{Z}$  can be decomposed into the sum of two matrices:  $\mathbf{Z}_{Orth}$  and  $\mathbf{Z}_{proj}$ . Namely:

$$\mathbf{Z} = \mathbf{Z}_{Orth} + \mathbf{Z}_{proj} \quad (10)$$

Where:

$\mathbf{Z}_{Orth}$  represents the  $\mathbf{Z}$  orthogonalized with respect to  $\mathbf{T}_X$ .

$\mathbf{Z}_{proj}$  is the projection of  $\mathbf{Z}$  on  $\mathbf{T}_X$  (and it is therefore contained in the column space of  $\mathbf{T}_X$ ).

From Eq. 10,  $\mathbf{Z}_{Orth} = \mathbf{Z} - \mathbf{Z}_{proj}$ , which rearranged explicitly for  $\mathbf{T}_X$  becomes:

$$\mathbf{Z}_{Orth} = \mathbf{Z} - \mathbf{T}_X (\mathbf{T}_X^T \mathbf{T}_X)^{-1} \mathbf{T}_X^T \mathbf{Z} \quad (11)$$

Mainly, orthogonalization “empties”  $\mathbf{Z}$  of its projection onto  $\mathbf{T}_X$  ( $\mathbf{Z}_{proj}$ ) removing the part that lies in the column space of  $\mathbf{T}_X$ .

This is a crucial step in the algorithm. In fact, orthogonalization allows performing the following step without any loss of information from  $\mathbf{Z}$ . Moreover, it grants to the method several benefits that will be exposed in detail in the following subparagraph.

**Step 3: Second Regression:**

In this case,  $\mathbf{E}$  from Eq. 9 is fitted to  $\mathbf{Z}_{orth}$  by PLS. Due to the orthogonalization, even if  $\mathbf{Z}_{proj}$  is not contributing to the fit, there is no loss of information. Since  $\mathbf{Z}_{proj}$  is in the column space of  $\mathbf{T}_X$ , it has already implicitly been modelled in the first fit.

In this step the  $\mathbf{Z}_{orth}$ -scores  $\mathbf{T}_{Z_{orth}}$ , the  $\mathbf{Z}_{orth}$ -loadings  $\mathbf{P}_{Z_{orth}}$ , the  $\mathbf{Z}_{orth}$ -weights  $\mathbf{W}_{Z_{orth}}$ , and the  $\mathbf{Y}$ -loadings  $\mathbf{Q}_{Z_{orth}}$  are obtained.

**Step 4: Predictive model:**

$\mathbf{T}_X$  and  $\mathbf{T}_{Z_{orth}}$  are orthogonal by construction. Consequently, the full predictive model can be calculated simply summing up the predictions from the individual regressions:

$$\hat{\mathbf{Y}} = \mathbf{T}_X \mathbf{Q}_X^T + \mathbf{T}_{Z_{orth}} \mathbf{Q}_{Z_{orth}}^T \quad (12)$$

If wanted, Eq.12 can be calculated using the original measures:

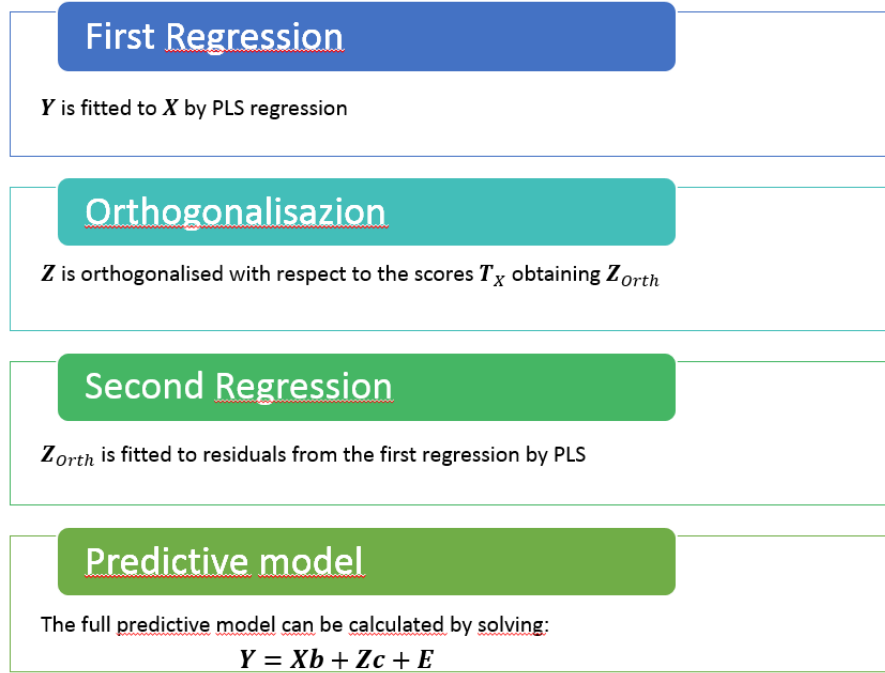
$$\hat{\mathbf{Y}} = \mathbf{X} \mathbf{V}_X \mathbf{Q}_X^T + \mathbf{Z}_{orth} \mathbf{V}_{Z_{orth}} \mathbf{Q}_{Z_{orth}}^T \quad (13)$$

Where  $\mathbf{V}_X (= \mathbf{W}_X (\mathbf{P}_X^T \mathbf{W}_X)^{-1})$  and  $\mathbf{V}_{Z_{orth}} (= \mathbf{W}_{Z_{orth}} (\mathbf{P}_{Z_{orth}}^T \mathbf{W}_{Z_{orth}})^{-1})$  are the weights allowing the direct calculation of the scores from the respective data blocks. Additionally, it is also possible to rearrange Eq.12 to be expressed in terms of  $\mathbf{Z}$  instead of  $\mathbf{Z}_{orth}$ :

$$\hat{\mathbf{Y}} = \mathbf{X} \mathbf{V}_X \mathbf{Q}_X^T + \mathbf{Z}_{orth} \mathbf{V}_{Z_{orth}} \mathbf{Q}_{Z_{orth}}^T \quad (14)$$

$$\hat{\mathbf{Y}} = \mathbf{X} \mathbf{V}_X \mathbf{Q}_X^T + (\mathbf{I} - \mathbf{T}_X (\mathbf{T}_X^T \mathbf{T}_X)^{-1} \mathbf{T}_X^T) \mathbf{Z} \mathbf{V}_{Z_{orth}} \mathbf{Q}_{Z_{orth}}^T \quad (15)$$

$$\hat{\mathbf{Y}} = \mathbf{X} \mathbf{V}_X (\mathbf{Q}_X^T - (\mathbf{T}_X^T \mathbf{T}_X)^{-1} \mathbf{T}_X^T \mathbf{Z} \mathbf{V}_{Z_{orth}} \mathbf{Q}_{Z_{orth}}^T) + \mathbf{Z} \mathbf{V}_{Z_{orth}} \mathbf{Q}_{Z_{orth}}^T \quad (16)$$



**Figure 4: Summary of SO-PLS's algorithm: The two predictor blocks case.**

SO-PLS is a sequential method, which, in principle, does not present any restriction regarding the number of the predictor blocks. When more than two predictor blocks are involved, any further block is orthogonalized with respect to the scores of all preceding blocks and then the full predictive model is calculated summing the contributions from the different regressions.

### Possible benefits of SO-PLS

SO-PLS provides a number of benefits that make it suitable both for prediction and interpretation purposes. The main ones are:

1. ***Problems related to ill-conditioned matrices are overcome.***

This is obviously due to the fact that PLS is involved. The reduction in latent variables solves problems related to ill-conditioned matrices. Note that this is in common with all the methods based on feature reduction.

2. ***SO-PLS is not affected by the blocks having different variances (scale invariance).***

Many multi-block methods require a preliminary scaling stage, to remove the spurious contributions related to the different variance of the matrices to be modeled. On the other hand, the characteristics of SO-PLS, i.e. sequential modeling carried out on blocks which are orthogonalized with respect to the scores extracted in the previous stages, make the method to be scale-invariant.

3. ***It allows the investigation of incremental contributions for each added block.***

Since SO-PLS involves the computation of as many PLS models as the number of predictor matrices, rather than a single, global, model, the contribution of each individual block can be more easily investigated. Moreover, due to the orthogonalization step, the contribution of each added block is additive (incremental) with respect to those already modeled. This means, for instance, that the inspection of the second PLS model can allow evaluating the effects of the addition of  $\mathbf{Z}$  and even assessing whether the inclusion of that particular block is worth or not. In this context, it should be highlighted that, since the interpretation of the second PLS model is slightly different from the interpretation of the first one, a detailed description of how to proceed is given in Paragraph 2.3.3.2.

4. ***The number of components for each PLS in the model can be defined for each block (independently on the others).***

Information is extracted from the predictors in as many different PLS regressions as the number of matrices. This means that the number of components to be used in each regression is optimized for that specific block.

Obviously, SO-PLS presents also some disadvantages. It is a quite young method; there are not many applications in literature and many aspects have not been explored in depth (in particular regarding the interpretation of the models). Another disadvantage is tightly related to the possibility of choosing the number of component for each block. In fact, despite this can be considered a benefit (point 4 above), it implies the optimization of a number of model parameters equal to the number of the blocks. Increasing the number of the optimized parameters, even the chance of overfitting could increase. Moreover, the selection of components results more complicated if compared to methods that apply the same complexity to all the predictor blocks (e.g. MB-PLS).

### **2.3.3.1 Estimating the optimal number of latent variables in SO-PLS**

There are at least two approaches to define the optimal number of latent variables in SO-PLS: the *global strategy* and the *sequential strategy*. In the first one, all the possible combinations of components (up to a predetermined maximum value) are tested. Namely, SO-PLS models are built with all the possible combinations of latent variables in the different blocks, and validated by cross-validation. The combination of components to be used will be selected based on the values of RMSECV (more details at the end of this paragraph). In the sequential strategy, the number of latent variables to be used is separately optimized for each regression



step: at first, the best number of latent variables for the regression model between  $\mathbf{X}$  and  $\mathbf{Y}$  is chosen, and only successively, the number of components for fitting the  $\mathbf{Y}$ -residuals to  $\mathbf{Z}_{Orth}$  is separately optimized. Also in this case, the decision is made on the basis of the RMSECV. Applying this strategy the order of the blocks is highly relevant.

It has to be stressed that, in both approaches, the combination of components giving the lowest RMSECV is not always the best solution. In fact, in this way the number of latent variables could be overestimated, leading to overfitted models. It is possible to impose parsimony testing the differences between the different RMSECVs. This is particularly relevant when it is not possible to visually inspect the results (i.e., in simulations). A suggested procedure to automatize the choice of the optimal complexity (applied in all the simulations presented in this work) is described at the end of this paragraph.

In order to understand the difference between the two approaches for selecting the number of components, a simulation study has been conducted. The study is divided in two parts, which reflect the different structures of the analyzed data sets. In both cases, data sets were built simulating two blocks of predictors and a response vector; moreover, for a proper validation of the obtained results, both training and test sets were generated. In order not to be biased by chance results, simulations have been replicated one hundred times.

In the first part of the simulation study, both predictor blocks were conceived to mimic spectral data. In the second part, one predictor block is spectra-like, while the other is built to mimic a matrix of process variables. More details about the simulations can be found in text box 1.

Both global and sequential strategies were applied to define the optimal complexity for SO-PLS models. Then, SO-PLS models were built on the training sets and validated on the external test sets, which were, then, completely independent, as they were not involved in the definition of the optimal complexity. When the optimal number of latent variables for the  $\mathbf{X}$  and the  $\mathbf{Z}$  blocks evaluated with the two approaches are the same, the models also coincide. The average values (over the one hundred replicates) of RMSECV and RMSEP are reported in Table 1. Additionally, the distribution of optimal number of latent variables is displayed in Figure 5. These results have constituted the basis for the evaluation of the two approaches.

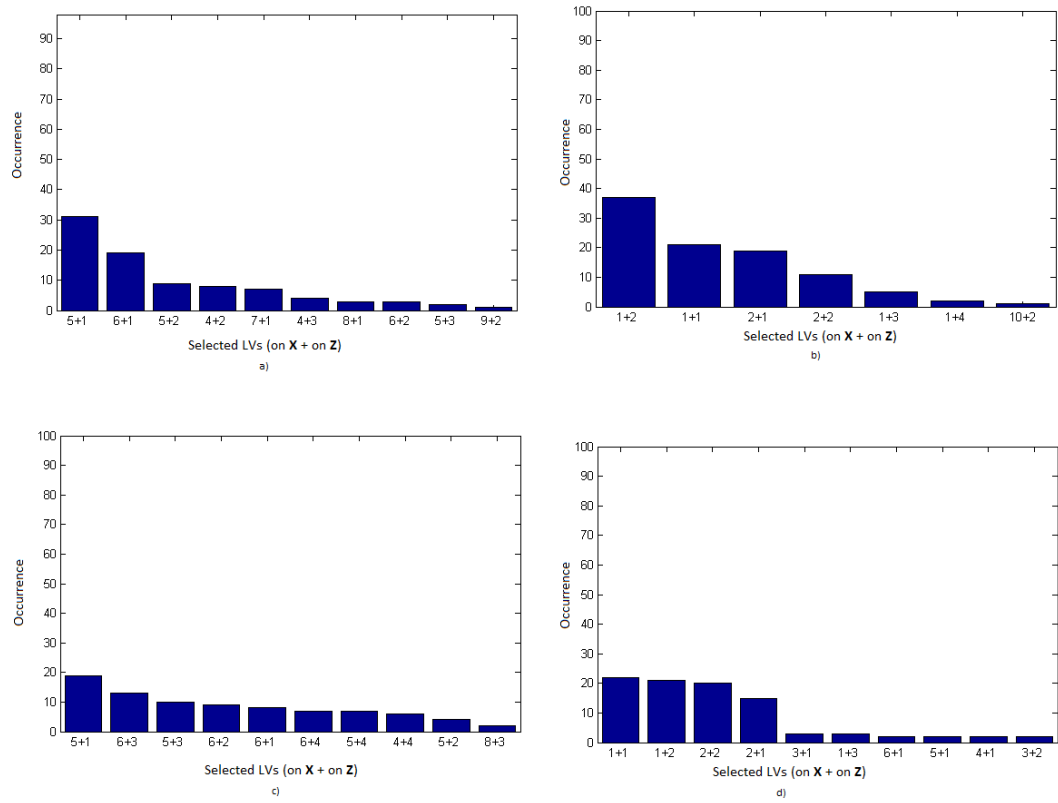
*Table 1: Averaged values (over the one hundred replicates) of RMSECVs and RMSEPs for the simulation study*

	<b>Simulation study – Part I</b>			
	<b>RMSECV</b>	<b>Std. Deviation</b>	<b>RMSEP</b>	<b>Std. Deviation</b>
<b>Sequential</b>	0.3375	0.1630	1.0504	0.5250
<b>Global</b>	0.3245	0.0436	0.3479	0.0261
	<b>Simulation study – Part II</b>			
	<b>RMSECV</b>	<b>Std. Deviation</b>	<b>RMSEP</b>	<b>Std. Deviation</b>
<b>Sequential</b>	0.3368	0.1661	1.1678	0.4912
<b>Global</b>	0.3186	0.0502	0.3230	0.0291

From the table, it is evident that, in this context, the two approaches are leading to different results. Independently of the data set structure, the sequential approach results in higher RMSECV and RMSEP.

Another important parameter to look at is the number of latent variables selected by each approach. The optimal complexity of the blocks is known a priori from the simulation, and it is five components for **X** and two for **Z** (one of which is common with **X**). Histograms in Figure 5 report the number of latent variables suggested by the two different approaches and their occurrence over the replicate models. Due to the common component, the “expected” optimal complexity is either five plus one or four plus two (for **X** and **Z**, respectively). As shown in the figure, especially for the first study, the global approach leads to the selection of the “expected” number of latent variables more frequently than the sequential one. Moreover, the sequential approach more often underestimates the number of components. This is probably the reason of the observed differences between the RMSECV and RMSEP obtained by this approach. All these considerations suggest that the global approach is more appropriate in identifying the optimal complexity in the data blocks.

This study was carried out just as a preliminary investigation in order to choose the approach to follow across all the research; therefore, it is not presented in any paper enclosed to this thesis.



**Figure 5: Latent variables selected in the Simulations Study (Text box 1):** a) using global approach in Simulation Study-Part I; b) using sequential approach in Simulation Study-Part I; c) using global approach in Simulation Study-Part II; d) using sequential approach in Simulation Study-Part II.

**Simulation Study-Part I**

$\mathbf{X}$  and  $\mathbf{Z}$  are spectra-like block,  $\mathbf{X}$ - and  $\mathbf{Z}$ -loadings are simulated as sum of Gaussians. In order to mimic a real data set, the two blocks are simulated having unique components  $\mathbf{T}_U$  (specific for each block) and a common component  $\mathbf{T}_C$ . The  $\mathbf{X}$ -block has four unique components plus the common one, while  $\mathbf{Z}$ -block presents one unique and one common component.  $\mathbf{X}$ -scores  $\mathbf{T}_X$  and  $\mathbf{Z}$ -scores  $\mathbf{T}_Z$  are generated as:

$$\mathbf{T}_X = [\mathbf{T}_C \mathbf{T}_{UX}] \text{ and } \mathbf{T}_Z = [\mathbf{T}_C \mathbf{T}_{UZ}] \quad (\text{Tb1, Tb2})$$

Where  $\mathbf{T}_{UX(N \times 4)}$  and  $\mathbf{T}_{UZ(N \times 1)}$  are the unique components for  $\mathbf{X}$  and  $\mathbf{Z}$ , respectively. All the scores are simulated from the normal distribution  $N(0,1)$ .  $\mathbf{X}$  and  $\mathbf{Z}$  are generated as a TP-product.  $\mathbf{X}$ - and  $\mathbf{Z}$ -loadings ( $\mathbf{P}_X$  and  $\mathbf{P}_Z$ ) are simulated as sum of Gaussians.  $\mathbf{X}$  and  $\mathbf{Z}$  dimensions are 100x500 and 100x300, respectively.

Finally,  $\mathbf{y}$  is generated as:  $\mathbf{y} = \mathbf{T}\mathbf{b}$ .  $\mathbf{T}$  are the concatenated scores from both blocks ( $\mathbf{T} = [\mathbf{T}_X \mathbf{T}_Z]$ ) while  $\mathbf{b}_{(7 \times 1)}$  is the coefficient vector generated as a matrix containing random values drawn from the uniform distribution in the open interval (0.05, 1.05).

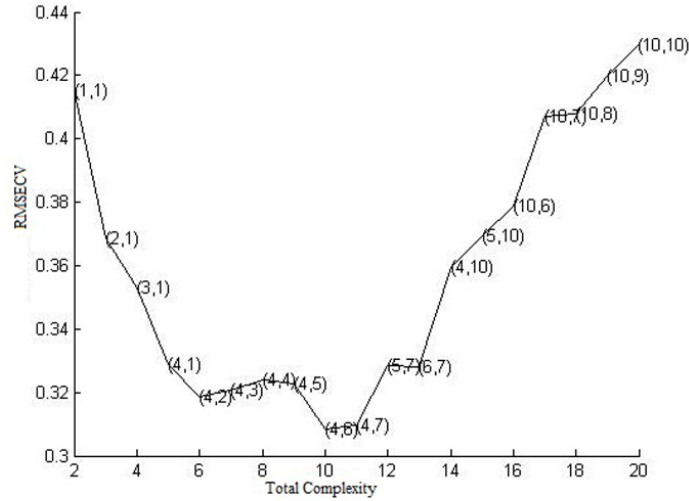
**Simulation Study-Part II**

$\mathbf{X}$  is spectra-like, and  $\mathbf{Z}$  mimics a process-block. Consequently,  $\mathbf{X}$ -loadings are generated as in Part I while  $\mathbf{Z}$ -loadings are simulated in order to reproduce process variables. A random number (between 1 or 2) of variables per component have a value between 0.75 and 1.25 while all the others have values between -0.25 and 0.25.  $\mathbf{X}$  and  $\mathbf{Z}$ ,  $\mathbf{T}$  and  $\mathbf{y}$  are built as in Part I (same number of components and same distinction between common and unique).  $\mathbf{X}$  and  $\mathbf{Z}$  have dimensions 100x500 and 100x15, respectively.

Accordingly, the global approach is the one followed in all the SO-PLS models presented in this work.

In the global strategy, RMSECV is calculated as a function of all the possible combinations of latent variables. Selection of the combination that gives the lowest error can be accomplished directly inspecting the numerical values obtained. It is also possible to make a graphical visualization of all the values, achieving a more effective and clear idea of which combination of latent variables is the most appropriate. These graphs (the so-called the *Måge plots*) are obtained plotting the RMSECVs as a function of the total number of component tested. An example is shown in Figure 6, where a maximum number of ten latent variables per block was allowed. In the figure, the two numbers in brackets indicate the components of the first and the second PLS regression, respectively. Note that, different combinations of latent variables present the same total number of latent variables (e.g., a total complexity of six results from combinations: one plus five, two plus four, three plus three, four plus two and five plus one). In order to obtain a clearer visualization, the combination of latent variables reported in brackets is the one (among all the combinations with the same total complexity) resulting in the lowest RMSECV (e.g., for a total number of latent variables equal to six, only four plus two is reported). This representation leads to two important considerations:

- 1) **Different combinations of latent variables, which sum up to the same *total amount* of latent variables (the entity called *total complexity* in the plot in Figure 6), can give different RMSECVs.** Different models having the same total complexity are built using different numbers of latent variables per block. Consequently, some models are built underestimating the number of components in one block and overestimating it in the other; predictions can be negatively affected by this.
- 2) **Combinations of latent variables that lead to models with different *total complexity* could give similar RMSECVs.** Looking at Figure 6, this would be the case of combinations four plus two and four plus four. They give approximately the same RMSECV. For the sake of parsimony, when two RMSECVs are not statistically different, the combination of latent variables suggested is the one with the smallest total complexity.



**Figure 6: Måge Plot for SO-PLS.** Number of latent variables per X- and Z-block, are reported in brackets in the plot.

Once the plot is created, it is quite straightforward to find the optimal combination of latent variables for the model. It has to be highlighted that the same plot can be prepared for any kind of regression model. This could be very useful to compare different (individual or multi-block) methods. A plot reporting a comparison between RMSECVs from different regression methods is shown in Figure 7 in Paper I.

As mentioned, sometimes it is not possible to inspect Måge plots (or RMSECVs values). In simulations the choice of the optimal combination of components needs to be automatic. One option would be to simply select the combination of latent variables which result in the lowest RMSECV. Nevertheless, this could lead to overestimation of the number of components and consequently to overfitted models. Consequently, in all the simulations presented in this work, a more parsimonious automatic selection of the optimal complexity was applied. Namely, the selected combination of components is the smallest one resulting in an RMSECV not statistically different from the absolute minimum of the error curve. Differences are tested by a  $\chi^2$  test (significance level 5%) [35].

### 2.3.3.2 Interpretation of SO-PLS models

Regarding interpretation, SO-PLS provides different possibilities. In particular, the effect of the addition of each block can be evaluated and interpreted.

### ***Scores Plots***

From SO-PLS models, interpretable *scores plot* can be obtained. In particular, together with the “usual” scores plots of the individual blocks, additional information can be obtained by graphing  $\mathbf{T}_X$  against  $\mathbf{T}_{Zorth}$ . Scores plot for SO-PLS can be found in *Paper I*. These plots can be interpreted as it is usually done for scores plots from PLS.

### ***Loadings Plots***

Loading plots can be created for each block involved in the SO-PLS model. These indicate which block-variables are influencing each component. In particular, in case of blocks of analytical measures, loadings maintain the chemical information. Interpretation of loadings from SO-PLS models is one of the focus in Paper IV. Simulation studies have been conducted to understand whether their inspection could lead to reasonable (and reliable) interpretation: it was found that loadings can be a relevant tool to inspect an SO-PLS model. Often, it came out that there is a slight difference between the “expected” number of components (ground truth) and the ones identified as interpretable according to the proposed *explained variance criterion* (Paper IV). This could indicate that SO-PLS models sometimes need one additional component to get rid of the noise. Nevertheless, this does not represent a huge overestimation of the latent variables.

Note that the procedure to display  $\mathbf{X}$ -loadings is different from the procedure used to plot  $\mathbf{Z}$ -loadings.  $\mathbf{X}$ -loadings plots can be displayed simply plotting  $\mathbf{P}_X$  loadings from the first regression of the SO-PLS model. Instead,  $\mathbf{Z}$ -loadings cannot be represented directly. In fact, due to the orthogonalization step,  $\mathbf{Z}$ -loadings may not be in the same row space of  $\mathbf{Z}$  (but they are in the same column space). They can be projected back in the original space before being interpreted. This is done calculating:

$$\mathbf{P}_Z = (\mathbf{T}_{Zorth}^T \mathbf{T}_{Zorth})^{-1} \mathbf{T}_{Zorth}^T \mathbf{Z} \quad (17)$$

### ***Regression coefficients Plot***

Regression coefficients are the coefficients of the linear combination(s) relating the predictors to the response(s). Their magnitude and sign may reflect the entity of the contribution of individual predictors to the  $\mathbf{Y}$ -variable(s). Accordingly, plotting the regression vector(s) can represent a further interpretation tool.

However, it has to be stressed that particular care should be taken when interpreting the regression coefficients. First of all, in order for them to truly reflect the relative importance of the predictors, proper scaling of the variable should be adopted. Moreover, the presence of interferences (whose signals overlap with those of the analyte(s) of interest) leads to regression coefficients showing profiles different from those of the pure substances. A wider discussion of these problems can be found in [36-37].

### 2.3.4 SO-N-PLS in detail

#### *A novel multi-block and multi-way method*

SO-N-PLS is a novel method proposed to build regression (and classification) models handling multi-way arrays (Paper III). This method is a natural extension of SO-PLS to the multi-way field.

In principle, multi-block methods can handle multi-way arrays after unfolding, but this procedure could lead to some issues; therefore, methods conceived for multi-way arrays are an actual need. As said in Paragraph 2.2, in literature there are some methods which allow the combination of blocks with a different number of modes. Despite this, this field is still relatively unexplored. Consequently, the possibility of having a multi-block method able to handle multi-way arrays and, at the same time, which retains the above exposed benefits of SO-PLS, was investigated. As a consequence, a novel method called SO-N-PLS was developed. This method is conceived to handle any kind of multi-way arrays; in this work, the focus has been on two- and three-way arrays. The method has been studied and tested using no more than two blocks of predictors at a time. Theoretically, it can be applied on a larger number of predictor arrays. The algorithm is, as for SO-PLS, divided in four steps (upon centering and possibly scaling of the data). The main difference between SO-PLS and SO-N-PLS is that PLS regression is replaced by N-PLS regression. Considering the case of two three-way predictors  $\underline{\mathbf{X}}$  and  $\underline{\mathbf{Z}}$ , the algorithm can be summarized as follows:

- 1)  $\mathbf{Y}$  is fitted to  $\underline{\mathbf{X}}$  by N-PLS regression.
- 2)  $\underline{\mathbf{Z}}$  is orthogonalized with respect to the scores ( $\mathbf{T}_X$ ) of the previous regression, obtaining  $\mathbf{Z}_{Orth}$ .
- 3) Residuals from step 1) are fitted to  $\mathbf{Z}_{Orth}$  by N-PLS regression.
- 4) The full predictive model can be computed from the equation:



$$\mathbf{Y} = \mathbf{X}_{un}\boldsymbol{\gamma} + \mathbf{Z}_{un}\mathbf{v} + \mathbf{E} \quad (18)$$

( $\boldsymbol{\gamma}_{(JK \times A)}$  and  $\mathbf{v}_{(HL \times A)}$  are the regression coefficients and  $\mathbf{E}_{(N \times A)}$  is the residual matrix)

In each step of the algorithm, some aspects need to be stressed:

### **Step 1**

N-PLS is a direct extension of classical PLS for N-dimensional data arrays [38-39]. In this specific case (three-way arrays), it is called tri-PLS, and the predictor array is decomposed by a tri-linear decomposition. Given  $\underline{\mathbf{X}}_{(N \times J \times K)}$ , the  $F$ -component model can be expressed as :

$$x_{njk} = \sum_{f=1}^F t_{nf} w_{jf}^J w_{kf}^K + e_{njk} \quad (19)$$

Where  $\mathbf{t}$  are the  $\mathbf{X}$ -scores and  $\mathbf{w}^J$  and  $\mathbf{w}^K$  are the  $\mathbf{X}$ -weights of the second and of the third mode, respectively. As in Martens' PLS algorithm [40], N-PLS does not present any additional sets of loading vectors  $\mathbf{p}$ . Components are extracted sequentially and loading weights  $\mathbf{w}$  provide scores having maximum covariance with the still unexplained part of  $\mathbf{Y}$ . These are different from the weights that would be extracted by PLS on a two-mode matrix or on an unfolded three-way matrix.

### **Step 2**

The orthogonalization step is performed on the unfolded three-way block.  $\mathbf{Z}_{orth}$  is obtained replacing  $\mathbf{Z}$  with  $\mathbf{Z}_{un}$  in Eq. 11. Finally,  $\mathbf{Z}_{orth}$  is refolded back to the original structure.

### **Step 3**

Once again regression is performed by N-PLS. Residuals from the first regression are fitted to  $\mathbf{Z}_{orth}$ . Comments made for step 1 apply also here.

### **Step 4**

Finally, the predictive model can be calculated. A regression equation function of the original variables (instead of score vectors), can be formulated in terms of unfolded matrices (Eq.18).

The equation can then be used for prediction of new samples.

Note that regression coefficients from SO-PLS (on unfolded blocks) and SO-N-PLS models have same size, but they are calculated differently. In the case of SO-N-PLS, regression coefficients are calculated following *Method 2* suggested in De Jong [41]. A weight matrix  $\mathbf{W}$  is calculated as:

$$\mathbf{W} = [\mathbf{w}_1^J \otimes \mathbf{w}_1^K \mathbf{w}_2^J \otimes \mathbf{w}_2^K \dots \mathbf{w}_F^J \otimes \mathbf{w}_F^K] \quad (20)$$

Then, the weights  $\mathbf{R}$ , allowing the direct calculation of the scores from the unfolded three-way array, are obtained as:

$$\mathbf{R} = \mathbf{W}/\delta \quad (21)$$

Where  $\delta$  is the upper triangular part of the inner product  $\mathbf{W}^T \mathbf{W}$ .

Finally, regression coefficients  $\mathbf{b}_{N-PLS}$  are calculated as:

$$\mathbf{b}_{N-PLS} = \mathbf{RQ}^T \quad (22)$$

### ***Benefits of SO-N-PLS***

SO-N-PLS has demonstrated to possess different benefits, both from the prediction and the interpretation point of view:

1. **Unfolding is not required:** SO-N-PLS allows handling multi-way data without performing any preliminary unfolding. This represents a huge advantage, since models built on unfolded data are more prone to overfitting and therefore less reliable. Moreover, being built on the original multi-way data, models result simpler and, in particular, info are condensed. Therefore, they lead to more effective interpretations.
2. **Good predictions on small noisy data sets:** When data have a clear three-way structure and there is linear relationship between predictors and response, SO-N-PLS gives accurate predictions. In particular, it performs better than MB-PLS and SO-PLS on highly noisy data sets. This indicates that SO-N-PLS filters out the noise better than the other two methods.
3. **SO-N-PLS finds the actual underlined complexity in data:** SO-N-PLS has demonstrated to select the actual complexity in data. This is an advantage from the interpretation point of view; in fact, more components would mean more parameters to interpret. Instead, selecting the smallest reasonable number of latent variables, information is condensed in the simplest model possible.
4. **Graphical interpretation:** A number of plots can be created to graphically interpret SO-N-PLS models. Scores, weights (outer product of  $\mathbf{w}^J$  and  $\mathbf{w}^K$ ) and regression coefficients can be plotted and interpreted (More details in the following paragraph).

### 2.3.4.1 Interpretation of SO-N-PLS models

#### *Scores Plots*

Scores plots in SO-N-PLS are displayed as in SO-PLS; they can be interpreted in the same way.

#### *Weights Plot*

Weights plots can be obtained in two different ways:

1. Plotting  $\mathbf{w}^J$  and  $\mathbf{w}^K$  individually
2. Plotting the outer product  $\mathbf{w}^J(\mathbf{w}^K)^T$  as a landscape

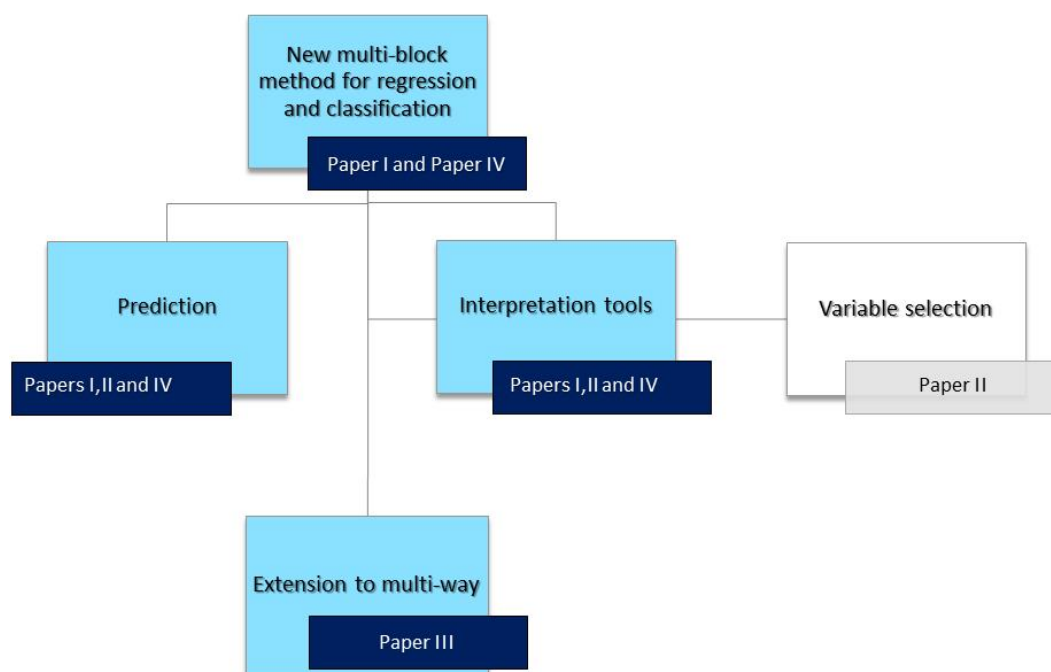
$\mathbf{X}$ -weights can be interpreted directly;  $\mathbf{Z}$ -weights have the same geometrical issue exposed above for  $\mathbf{Z}$ -loadings (see Paragraph 2.3.3.2). Therefore, they also need to be projected back in the  $\mathbf{Z}$  space:

$$\mathbf{W}_z = (\mathbf{T}_{\text{Zorth}}^T \mathbf{T}_{\text{Zorth}})^{-1} \mathbf{T}_{\text{Zorth}}^T \mathbf{Z}_{\text{un}} \quad (23)$$

#### *Regression coefficients Plot*

Regression coefficients ( $\boldsymbol{\gamma}$  and  $\mathbf{v}$  in Eq.18) can be reshaped and plotted as landscapes. For the interpretation, the same consideration exposed in Paragraph 2.3.3.2 apply.

## Chapter 3: Multi-block Classification



**Figure 7: Conceptual Flow chart: Multi-block classification.** Models are used to predict class-belonging and to interpret the system.

### 3.1 Brief introduction to classification

In data analysis, it can be useful to group objects on the base of some specific characteristics they have. Objects with similar features will belong to the same *class* or *category*.

Consequently, the term *classification* refers to the capability of separating objects depending on specific characteristics. These characteristics can be measured, or known *a priori*. For instance, samples can be classified on the basis of the content of a specific compound, or because of their geographical origin. Obviously, changing the qualitative information one is looking at, the same objects will give rise to different groupings.

In any case, in order to assign objects to categories, a classification model is needed. Once the calibration model is built, it is possible to predict class-belonging for unknown samples.

In literature, there are plenty of classification methods. They are mainly divided in methods that focus on modelling the boundaries among categories (*discriminant classification*) and those which focus on identifying the portion of space occupied by a specific class (*class-modelling classification*) [42]. A brief description of the two is given below.

### *Discriminant methods*

Applying discriminant methods, classification is based on differences between samples coming from the various categories. These methods divide the space of the variables in as many regions as the number of the groups in the calibration set. Consequently, each sample is assigned to a specific category.

These methods define the class-belonging on the base of the Bayes' rule [43]. Therefore, the (*posterior*) probability that each sample belong to a class is calculated for all the classes. Then, it is assigned to the class with the highest probability. Consequently, what mainly differ these methods, is the way this probability is calculated.

Given a sample  $\mathbf{x}_i$ , its posterior probability  $p(g|\mathbf{x}_i)$  of belonging to the  $g$ -th class can be expressed as:

$$p(g|\mathbf{x}_i) \propto p(\mathbf{x}_i|g)p_0(g) \quad (24)$$

Where  $p(\mathbf{x}_i|g)$  is the likelihood and  $p_0(g)$  is the a priori probability that a sample belong to the specific class  $g$ .

### *Class-modelling classification methods*

Class-modelling classification methods are many, and quite different from each other [44-46]. Applying this classification procedure, the attention is more on *intra*class similarities, than on *inter*class differences (as it is for discriminant methods). Consequently, these methods focus on specific characteristics of each group. Since each category is modelled independently on the others, not all the variable space corresponds to class-belonging regions. Therefore, it is possible that some samples are not assigned to any category (or they can pertain to more than one group).

Additionally, classification methods can also be differentiated because of the mathematical nature of the classification-rule. This can be *linear* or *non-linear*.

In the present work, the attention has been on discriminant classification. This family of methods have been preferred because of the univocity of the response. In particular, the focus has been on the Linear Discriminant Analysis (LDA) by Fisher [47].

## **3.2 Linear Discriminant Analysis (LDA)**

The Linear Discriminant Analysis (LDA) is a method proposed by Fisher [47], and it is one of the first of this genre. LDA's target is to find the linear surfaces that optimize the separation between the different class-regions of the system under study [48-49].

The method relies on two main assumptions:

1. Samples are normally distributed (among each class)
2. The dispersion is the same in each class

From the first assumption, for the  $g$ -th class, it follows that:

$$p(\mathbf{x}_n|g) = \frac{1}{(2\pi)^{\frac{J}{2}}|\mathbf{S}_g|} e^{-\frac{1}{2}(\mathbf{x}_n - \bar{\mathbf{x}}_g)^T \mathbf{S}_g^{-1}(\mathbf{x}_n - \bar{\mathbf{x}}_g)} \quad (25)$$

where:

$\bar{\mathbf{x}}_g$  and  $\mathbf{S}_g$  are the centroid and the variance/covariance matrix for class  $g$ , respectively, while  $J$  is the number of variables.

From the second assertion, it comes that the variance/covariance matrix is the same for all the categories and it is:

$$\mathbf{S} = \frac{\sum_{g=1}^G (n_g - 1) \mathbf{S}_g}{N - G} \quad (26)$$

Where  $N$ ,  $n_g$  and  $G$  are the total number of samples, the number of samples in the  $g$ -th class and the total number of classes, respectively.

The probability that a sample belongs to a specific  $g$  class corresponds to:

$$p(g|\mathbf{x}_n) = \frac{c_g p_0(g)}{(2\pi)^{\frac{J}{2}}|\mathbf{S}|} e^{-\frac{1}{2}(\mathbf{x}_n - \bar{\mathbf{x}}_g)^T \mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}}_g)} \quad (27)$$

Where  $c_g$  is a normalization factor which takes into account the requirement that the sum of the probabilities that a sample belongs to each category is 1.

In agreement with the Bayes' rule, a sample belonging to class  $I$  will fall in the region defined by:

$$p(I|\mathbf{x}) > p(M|\mathbf{x}) \quad \forall M = 1 \dots G, M \neq I \quad (28)$$

Consequently, for more than two categories, more than one hyperplane is needed to divide the  $G$  classes. Each pair of classes is divided by a decision boundary. The boundary correspond to that portion of the space where the probability that a sample belongs to each class of the pair is the same. Mathematically,  $p(I|\mathbf{x}) = p(M|\mathbf{x})$ .

Which can be written:

$$\log\left(\frac{c_g p_0(I)}{c_g p_0(M)}\right) - \frac{1}{2}(\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_M)^T \mathbf{S}^{-1}(\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_M) + (\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_M)^T \mathbf{S}^{-1} \mathbf{x} = 0 \quad (29)$$

$$\forall I, M = 1 \dots G, I \neq M$$

Defining  $w_0 = \log \left( \frac{c_g p_0(I)}{c_g p_0(M)} \right) - \frac{1}{2} (\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_M)^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_M)$  and  $w_T = (\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_M)^T \mathbf{S}^{-1}$ , Eq.29 becomes:

$$w_0 + \mathbf{w}_T \mathbf{x} = 0 \quad (30)$$

Equation 30 suggests that the data can be projected onto a direction orthogonal to the decision boundaries ( $\mathbf{w}_T$ ). Such direction is called a canonical variate, and the corresponding scores  $C$  are defined by  $w_0 + \mathbf{w}_T \mathbf{x} = C$ . The canonical variate represents the direction (orthogonal to the decision boundary, i.e. the hyper-plane represented by Eq. 30) of maximum discrimination between the classes. It maximizes the ratio between the between-class variance (the covariance between the samples of different classes, i.e. the distance between the centroids) and the within-class covariance (the covariance between samples in the same class) [34, 50-52].

In the multi-category case, the between variance/covariance matrix  $\mathbf{S}_b$  can be defined as:

$$\mathbf{S}_b = \frac{1}{G} \sum_{g=1}^G (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^T \quad (31)$$

Where  $\bar{\mathbf{x}}_g$  and  $\bar{\mathbf{x}}$  are the mean vectors of class  $g$  and of all the samples, respectively.

In this case, the canonical variates that maximize the separation among classes are the eigenvectors  $\mathbf{w}_f$  corresponding to the largest  $\lambda_f$  in :

$$\mathbf{S}^{-1} \mathbf{S}_b \mathbf{w}_f = \lambda_f \mathbf{w}_f \quad (32)$$

This means canonical variates are the eigenvectors which maximize the ratio  $\frac{S_b}{S}$  i.e. which simultaneously maximize the distance between the centroids of the groups and minimize the distances between samples within each group. This makes the canonical variates particularly relevant for the representation of LDA outcomes. A graphical tool conceived for the inspection of classification results based on projection of samples in the space of canonical variates has been presented in Paper I and Paper III and exposed below in Paragraph 3.3.3. From Eq. 32 is evident that LDA requires that  $\mathbf{S}$  is an invertible matrix. This means, for example, that the number of objects should be at least equal to the number of variables and that the predictor are as least correlated as possible. These conditions do not occur often. In order to overcome the problem, different solutions have been proposed. One of these is to use PLS to obtain few and uncorrelated (latent) variables before applying LDA. This is discussed widely in the following paragraph.

### 3.3 PLS based classification methods

As said above, invertibility of the variance/covariance matrix  $\mathbf{S}$  is one of the constraints for the application of LDA. In literature, many of the proposed solutions involve the projection of the predictor matrix onto a relevant sub-space of latent variables by PLS [53-56].

#### 3.3.1 SO-PLS-LDA in detail

SO-PLS can be a starting point for a classification model. Paper I is focused on the extension of SO-PLS to the classification field. SO-PLS has been combined with LDA since linear discriminant analysis fits well with the sequential philosophy of SO-PLS. It has been demonstrated in Paper I that this approach gives good prediction ability and that it is easy to compute.

In this work, SO-PLS-LDA has been applied considering two predictor blocks  $\mathbf{X}$  and  $\mathbf{Z}$  and a categorical dummy matrix  $\mathbf{Y}$  (reporting the class information) but in principle this can be extended to even more blocks. This method has been tested on both simulated and real data sets obtaining good results. From the prediction point of view (See Paper I and Paper III), SO-PLS-LDA has demonstrated to be comparable to MB-PLS-LDA and SO-N-PLS-LDA (see Paragraph 3.3.2). Concerning interpretation, it leads to a number of graphical interpretation tools discussed below in Paragraph 3.3.3 and in Papers I and III.

SO-PLS-LDA algorithm can be summarized as follows:

1. The SO-PLS model is created as exposed in Paragraph 2.3.3.  $\mathbf{Y}$  is a categorical dummy matrix reporting the class information.
2.  $\mathbf{T}_X$  and  $\mathbf{T}_{Zorth}$  are concatenated obtaining a total score matrix  $\mathbf{T} = [\mathbf{T}_X \ \mathbf{T}_{Zorth}]$
3. LDA is applied to the scores  $\mathbf{T}$  [56].

As demonstrated in [56] LDA can safely be applied on scores instead of on predicted values.

##### 3.3.1.1 Estimation of the optimal number of latent variables in SO-PLS-LDA

The optimal complexity in SO-PLS-LDA is defined similarly as it is in SO-PLS (described in Paragraph 2.3.3.1). Also in this case, Måge plots can be exploited for picking the best combination of latent variables. In paper I two different ways of displaying the Måge plot for classification results are discussed. In fact, these plots can be obtained as explained in paragraph 2.3.3.1, or using the classification error instead of the RMSE. The classification



error fits better with the classification philosophy and is therefore the suggested approach. Nevertheless, when the number of samples per each class are few, it is better to use the RMSE. In fact, in this case, the classification error is an unstable measure of model quality. Classification error is a (more) crisp criterion while the dummy  $\mathbf{y}$  has continuous values. When there are only few samples per class, classification results will be affected by the specific sub-set used in each cross-validation loop. This is due to the fact that the misclassification of only one sample highly increase the classification error, magnifying the difference between two models that are actually giving similar classification rates. As a consequence, the resulting classification error could be a non-reliable parameter to define the optimal complexity to be used for the final model.

### 3.3.2 SO-N-PLS-LDA in detail

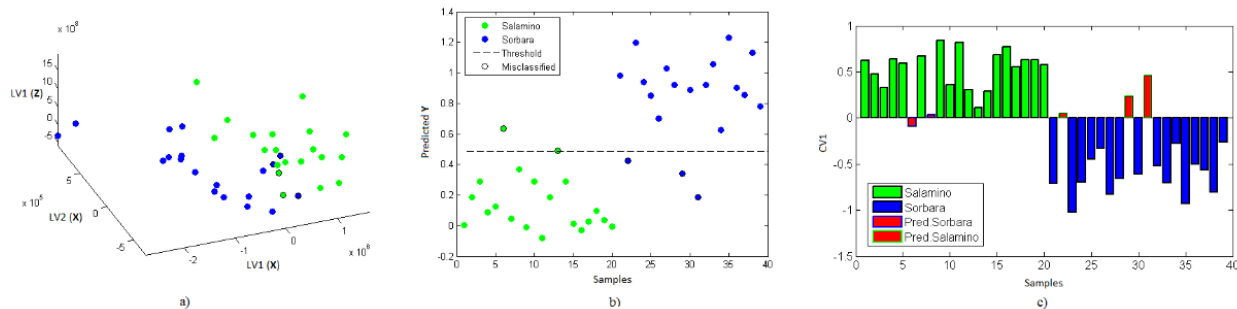
As for SO-PLS-LDA, LDA can be applied on the total scores from SO-N-PLS in order to create classification models. The number of components are chosen as in SO-PLS-LDA (Paragraph 3.3.1.1). Interpretation of results is discussed below.

### 3.3.3 Interpretation of classification results in SO-PLS-based models

Classification results are usually reported stating the amount of samples correctly classified (or the number of misclassified samples) out of the total amount. The so-called *classification error* corresponds to the ratio of the number of misclassified samples over the total number of samples. It can be reported as a decimal number or as a percentage.

Prediction is not the only goal in classification; interpretation of the results has a high relevance as well. Consequently, graphical representation of results is very relevant in the inspection of the system under study. PLS-LDA results are generally investigated looking at scores plots, regression coefficients plots or inspecting the  $\mathbf{Y}$ -predicted from the classification model. These representations (after small adaptations to the multi-block case) also suit SO-PLS-LDA models. Examples of scores plots are shown in Papers I and III and displayed in Figure 8a. A  $\mathbf{Y}$ -predicted plot is shown in Figure 8b. Moreover, an additional representation (based on canonical variates), which fit the SO-PLS-LDA philosophy, is proposed in Papers I and III and displayed in Figure 8c.

Figure 8 is based on SO-N-PLS models calculated on the *reduced Lambrusco data set* (for a detailed description of the data set and of the classification problem, see [57] and Paper III).



**Figure 8: Proposed graphical representation tools for SO-PLS-LDA models** (Reduced Lambrusco data set): Each class is represented by a type of Lambrusco wine: Lambrusco Salamino (in green) and Lambrusco Sorbara (in blue). Of these thirty nine samples (twenty for Salamino and nineteen for Sorbara), SO-N-PLS-LDA (but also SO-PLS and MB-PLS) misclassified five samples in total (two for Salamino and three for Sorbara). a) Scores Plot; b) Y-predicted plot; c) Canonical Variates plot.

### Scores Plots Interpretation

A straightforward way of inquiring classification results is looking at the scores plot. Looking at samples in the space of the latent variables it is easier to observe groupings or to spot “suspicious” samples (like outliers). In SO-(N)-PLS-LDA contributions from different blocks of predictors have to be accounted for at the same time. Samples can be projected in the space of the first  $\mathbf{X}$ - and  $\mathbf{Z}_{orth}$ -scores (as in Figure 5 in Paper I). In Figure 8a, samples are displayed in the space of the first two  $\mathbf{X}$ -components and the first  $\mathbf{Z}_{orth}$ -component; misclassified samples are circled in black. The model is built using two components for the  $\mathbf{X}$ -block and one for  $\mathbf{Z}$ . Even if samples are shown in the space of all these components, the two classes look overlapped. Consequently, the investigation of the scores plot would not be a particularly suitable tool for the interpretation of this system.

### Inspection of Predicted Y

A further interpretation of the classification results can be done plotting the  $\mathbf{Y}$  predicted from the classification model. This kind of interpretation results particularly useful for two-class problems. In fact, in this case, the threshold, used to define the class belonging, acts as a boundary which distinguish the two classes. An example of this representation is reported in Figure 8b. The class boundary is estimated by applying LDA directly on the predicted  $\mathbf{Y}$  vector. This plot gives a good representation of the results. The two groups appear separated except for the misclassified samples (circled in black).

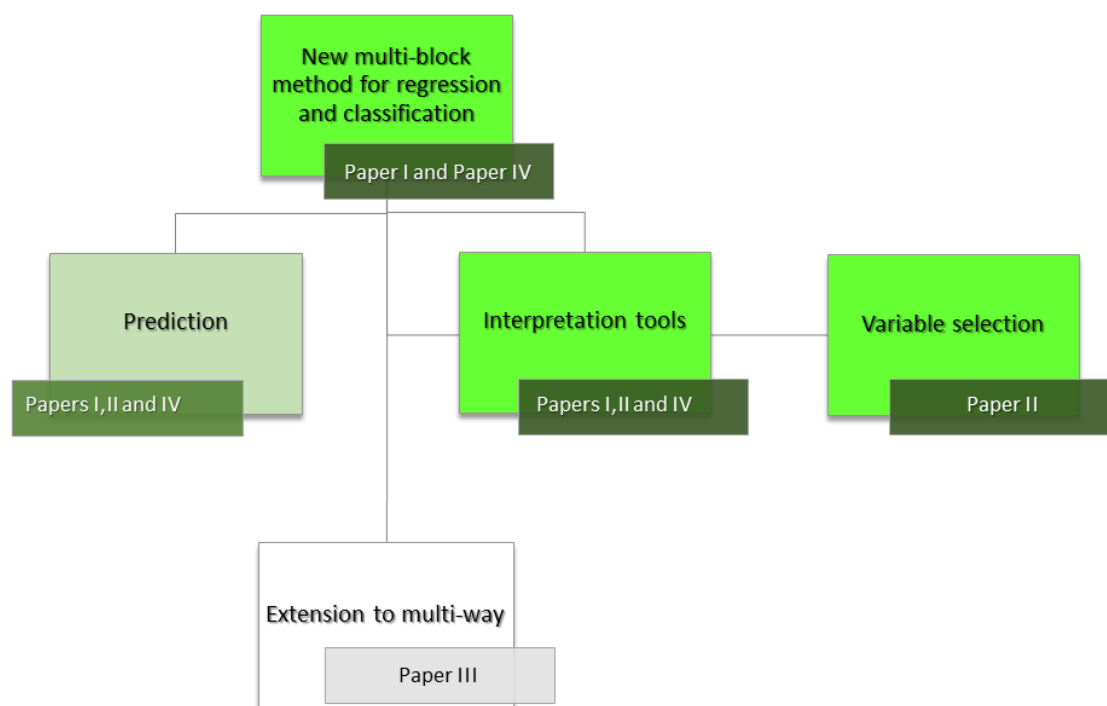
### ***Inspection of samples projected onto the canonical variates space***

An additional way of interpreting the classification results looking at samples in the space of the canonical variates is here proposed. This graphical tool is presented in Paper I and applied also in Paper III. As explained in Paragraph 3.2, canonical variates are intrinsically related to classification; consequently, their use for interpretation of the results is natural. A graphical tool for interpretation based on them has been developed. An example is displayed in Figure 8c. This plot is created using the cross-validated **Y**-values to calculate the covariance matrices used for the extraction of the canonical variates. For the reasons explained above in Paragraph 3.2, projecting samples onto this sub-space represents the most effective way of graphically highlighting the different grouping tendencies. In Figure 8c the two classes appear separated. *Salamino* samples have positive values, while *Sorbara* samples have negative ones. Misclassified samples are those which do not follow this trend (except for sample number eight, which has a very low but positive value).

### ***Interpretation of regression coefficients***

Outcomes from a PLS-based classification model can be interpreted examining the regression coefficients from the PLS (or N-PLS) involved. In fact, since the coefficient represent the relation between each block of predictors and the dummy matrix, they can provide information about which variable is more relevant for classification. Additionally, they can also indicate the “direction” of the discrimination, i.e., whether a predictor is higher (or lower) in a class than the other(s). However, it should be recalled that this type of interpretation is not so straightforward and not always completely reliable, as already discussed in Paragraph 2.3.3.2.

## Chapter 4: Variable Selection



**Figure 9: Conceptual Flow Chart: Variable selection.** Different tools for interpreting multi-block model can be obtained. Variable selection can be seen as one of them.

*Variable selection* is a procedure proposed in order to select a sub-set of variables to be used for the creation of a reduced regression model. It was originally introduced for handling stability problems in situations with high collinearity in data [58-59]. Often, removing non-informative and noisy variables generally improves prediction ability. Moreover, reducing variables, the system becomes simpler, with the consequence that it can be more easily interpreted. It has to be stressed that, since the selection of variables represent an additional model optimization step, validation is particularly relevant to avoid overfitted models [60]. Another important topic to be aware of is the presence of outliers. Variable selection relies on evaluating (often very little) differences in quality indices (e.g., RMSECV) or on the evaluation of significance model parameters. Consequently, the presence of outliers could mislead the selection of variables with the risk of ending up with a non-reliable sub-set of predictors.

Different categories of methods to select variables have been proposed in classical regression. The importance of the variables can be assessed through model parameters, diagnostic tools, classical statistical testing approaches or by a combination of these [61-66]. In some cases, (e.g. in spectroscopy or chromatography) it is more reasonable to selected interval of

contiguous variables instead of inspecting each individual predictor (e.g., as in iPLS and in related methods [67-68]).

Paper II is focused on the introduction of variable selection in a multi-block context. These two fields are seldom combined, and the literature lacks discussion on it. Some strategies for combining multi-block and variable selection are discussed in detail in Paragraph 4.2. The thesis is restricted to variable selection in the two methods: MB-PLS and SO-PLS.

In order to decide which standard variable selection method to use in the combination with multi-block regression, seven published variable selection methods have been taken into consideration. The three most promising have been selected: selectivity ratio, VIP and forward selection. This additional reduction has been based on an explorative simulation study described in Paper II. In particular, a *desirability index* ( $di$ ) has been developed in order to evaluate the suitability of the different approaches in this context. The idea behind the index is to find methods that select the best sub-set of variables from the interpretation point of view. This index is based on relative percentage of selected:

1. ‘selective’ variables ( $R_{sel}$ )
2. ‘irrelevant’ variables ( $R_{irr}$ )
3. ‘relevant but not selective’ variables ( $R_{rns}$ )
4. noise-variables ( $R_{noise}$ )

In this case, all of them are used as fractions between zero and one (for more details about these factors, please look at Appendix B in Paper II).

High values (close to one) of ‘selective’ variables and ‘relevant but not selective’ variables have a good influence on the final model while close to one values of ‘irrelevant’ and noise variables have a bad one. Consequently, the *desirability index* has been calculated as:

$$di = \sqrt[4]{R_{sel} \cdot (1 - R_{irr}) \cdot R_{rns} \cdot (1 - R_{noise})} \quad (33)$$

Another desirability index is suggested in Paper II. This is more prediction-oriented, meaning that it is conceived to figure out which variable selection method selects the sub-sets of variables leading to the best predictions. In fact, it takes into account the explained variance ( $Varsel$ ) and those factors deemed to mislead more the predictions, namely the relative

percentage of selected ‘irrelevant’ and noise variables. *Desirability index for predictions* ( $di_p$ ) is calculated as:

$$di_p = \sqrt[3]{Varsel \cdot (1 - Rirr) \cdot (1 - Rnoise)} \quad (34)$$

All the variable selection methods inspected are described in the following paragraph. The different procedures for variable selection in multi-block are mentioned in Paragraph 4.2 and described in detail in Paper II.

#### 4.1 Variable selection methods used in the work

Different variable selection methods have been used in this work. These are:

1. Selectivity Ratio (SR)
2. Variable Importance in Projection (VIP)
3. Significance multivariate correlation (sMC)
4. Elimination of Uninformative Variables for multivariate calibration (UVE)
5. Truncation-PLS (t-PLS)
6. Forward Selection
7. Jackknifing

Some of the methods rely on the evaluation of (regression) model parameters. Regression coefficients or loadings can indicate how variables are affecting the model. Usually, high values correspond to high relevance of the variable, while low ones indicate their irrelevance. Additionally, these model parameters can be used to generate further indices with the same aim, but that are supposed to be more accurate in same situations.

Other methods are based on classical statistics approaches, readapted in the light of chemometrics. A brief description of the listed methods is reported below.

##### *Selectivity Ratio (SR)*

Selectivity ratio (SR) [69-71] is the ratio between the variance explained by each predictor and the residual variance. For a  $\mathbf{X}$  matrix with  $N$  observations and  $J$  variables, it can be expressed like:

$$SR_j = \frac{\sum_{i=1}^N \hat{x}_{ij}^2}{\sum_{i=1}^N e_{ij}^2} \quad (35)$$

The approach followed is the one suggested by *Kvalheim* in [70] where it is shown that  $\mathbf{X}$  can be estimated by the means of the *Target Projection* (TP)-loadings ( $\mathbf{p}_{TP}$ ).

These  $\mathbf{p}_{TP}$  are considered a useful tool to interpret PLS-models since the corresponding TP-scores ( $\mathbf{t}_{TP}$ ) are proportional to the predictive response  $\hat{\mathbf{y}}$ .

TP-weights are defined as:

$$\mathbf{w}_{TP} = \frac{\mathbf{b}}{\|\mathbf{b}\|} \quad (36)$$

TP-scores ( $\mathbf{t}_{TP}$ ) can be calculated as:

$$\mathbf{t}_{TP} = \mathbf{X}\mathbf{w}_{TP} = \frac{\hat{\mathbf{y}}}{\|\mathbf{b}\|} \quad (37)$$

And TP-loadings ( $\mathbf{p}_{TP}$ ) can be obtained simply projecting the  $\mathbf{X}$  on the  $\mathbf{t}_{TP}$ :

$$\mathbf{p}_{TP} = \frac{\mathbf{X}^T \mathbf{t}_{TP}}{\mathbf{t}_{TP}^T \mathbf{t}_{TP}} \quad (38)$$

At the end,  $\hat{\mathbf{X}}$  can be written as :

$$\hat{\mathbf{X}} = \mathbf{t}_{TP} \mathbf{p}_{TP}^T \quad (39)$$

The variance and the residuals can be estimated by Eq. 39, and then replaced in Eq. 35.

Originally [70], an F-distribution-based cut off was proposed to point out relevant variables. This value is the one that would be chosen testing the statistical difference between numerator and denominator in Eq. 35. If a  $SR_j$  is greater than the critical value of the F-distribution, the corresponding variable is considered significant and it is selected (F-test with fixed false-rejection probability at 0.05 and  $N-2$  and  $N-3$  degrees of freedom).

As discussed in [71], this cut off value is not always the most appropriate choice.

In fact, for a relatively high number of samples, the critical value of the F-distribution approaches one. This could lead to two opposite scenarios: Selectivity Ratio is a too parsimonious method, or plenty of variables are selected as relevant. A critical F-value close to one cuts off variables that are explaining around 50% of the total variance; which is actually not that low. If this occurs for few variables, the F-test appears as a proper cut off. On the contrary, for some data, it could result an extremely low threshold, nullifying the variable

selection intent. For this reason, an alternative cut off value (to be used if the F-test critical value presents the above exposed problem) has been proposed in Paper II.

This alternative cut off value is the mean of the  $SR_j$  values. This is a robust threshold, conceived to fit SR in its own range of distribution. In particular, in Paper II, the mean  $SR$  has been preferred as a cut off value for the multi-block simulations.

#### Variable Importance in Projection (VIP)

The *variable importance in projection* (VIP) [72-73] is another model-based method widely used to select features. It is an index of the significance of variables in defining the  $\mathbf{X}$ - and  $\mathbf{Y}$ -spaces in a PLS model. Mathematically, VIPs are defined as:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^F w_{jf}^2 \|\hat{\mathbf{Y}}_f\|^2 J}{\|\hat{\mathbf{Y}}\|^2_F}} \quad (40)$$

Where  $J$  is the number of predictors.

$w_{jf}$  is the weight of the  $j$ -th predictor for the  $f$ -th PLS-component.

$\hat{\mathbf{Y}}_f$  and  $\hat{\mathbf{Y}}$  are the matrix of responses predicted using the  $f$ -th component only and the full model, respectively.

$\hat{\mathbf{Y}}_f$  and  $\hat{\mathbf{Y}}$  are calculated, respectively, as:

$$\hat{\mathbf{Y}}_f = \mathbf{t}_f \mathbf{q}_f^T \text{ and } \hat{\mathbf{Y}} = \mathbf{T} \mathbf{Q}^T \quad (41,42)$$

Where:

$\mathbf{t}_f$  and  $\mathbf{T}$  are the scores for the  $f$ -th variable and the scores matrix of the regression, respectively.

$\mathbf{q}_f$  and  $\mathbf{Q}$  are the  $\mathbf{Y}$ -loadings for the  $f$ -th variable and the scores matrix of the regression, respectively.

Therefore, VIPs are a measure of how much of the variance of  $\mathbf{X}$  is explained by each variables and, at the same time, of the correlation of  $\mathbf{X}$ 's with  $\mathbf{Y}$ .

The average value of the index is, by construction, equal to one. Then, the contribution of those variables which result in a VIP bigger than one is considered relevant (and therefore those variables are selected).



### *Significance multivariate correlation (sMC)*

Significance multivariate correlation (sMC) estimates the sources of variability (for each variable) coming from a PLS-regression [74].

The starting point of significance multivariate correlation is Selectivity Ratio, but the estimation of  $\mathbf{X}$  is done dropping the calculation of the loadings. In fact, according to the authors, the loadings  $\mathbf{P}$  coincide with the TP-weights:

$$\mathbf{P}_{sMC} = \frac{\widehat{\mathbf{b}_{PLS}}}{\|\widehat{\mathbf{b}_{PLS}}\|} \quad (43)$$

Therefore:

$$\mathbf{T}_{sMC} = \mathbf{X}\mathbf{P}_{sMC} = \mathbf{X} \frac{\widehat{\mathbf{b}_{PLS}}}{\|\widehat{\mathbf{b}_{PLS}}\|} \quad (44)$$

That is equivalent to:

$$\mathbf{T}_{sMC} = \mathbf{X} \frac{\hat{\mathbf{y}}}{\|\widehat{\mathbf{b}_{PLS}}\|} \quad (45)$$

In the end,  $\mathbf{X}$  can be estimated as:

$$\mathbf{X} = \mathbf{T}_{sMC} \mathbf{P}_{sMC}^T + \mathbf{E}_{sMC} = \frac{\widehat{\mathbf{y}} \widehat{\mathbf{b}_{PLS}}^T}{\|\widehat{\mathbf{b}_{PLS}}\|^2} + \mathbf{E}_{sMC} \quad (46)$$

Then, in order to define which variables are relevant, the ratio between the (variable-wise) Mean Squared Error of the model (MSE) and its mean squared residuals are compared to an F-test with 1 and  $N - 2$  degree of freedom.

$$sMC_i = \frac{MS_{i_{PLSreg}}}{MS_{i_{PLSresidual}}} = \frac{\frac{SS_{i_{PLSreg}}}{1}}{\frac{SS_{i_{PLSresidual}}}{N-2}} = \frac{\frac{\|\widehat{\mathbf{y}} \widehat{\mathbf{b}_{PLS}}^T\|^2}{\|\widehat{\mathbf{b}_{PLS}}\|^2}}{\frac{\|x_i - \frac{\widehat{\mathbf{y}} \widehat{\mathbf{b}_{PLS}}^T}{\|\widehat{\mathbf{b}_{PLS}}\|^2}\|^2}{N-2}} \quad (47)$$

Variables that exceed the F-test threshold value are considered statistically relevant and then they will be selected.

### *Elimination of Uninformative Variables for multivariate calibration (UVE)*

The method is based on the analysis of the  $\mathbf{b}$  regression coefficients [75].

First of all  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  by PLS regression. Then, an artificial variable matrix  $\Phi$  (generated randomly and multiplied by a small constant like  $10^{-10}$ ) of the same dimension of  $\mathbf{X}$  is created. These two matrices are then concatenated resulting in the matrix  $\mathbf{X}\Phi$  of dimensions  $N \times 2J$ . Another PLS-model for  $\mathbf{X}\Phi$  is calculated using leave one out procedure and taking the same complexity of the previous PLS model. In this way,  $N$  estimates of the  $2J$  regression

coefficients are obtained and organized in a matrix **B** of dimensions  $N \times 2J$ . Then, the mean of each column of **B** and the standard deviation for each variable are calculated. Then the *reliability*  $c_j$  can be estimated as:

$$c_j = \frac{b_j}{s(b_j)} = \frac{\frac{\sum_{i=1}^N b_{ij}}{n}}{\sqrt{\frac{\sum_{i=1}^N (b_{ij} - b_j)^2}{n-1}}} \quad \text{for } j = 1, \dots, J \quad (48)$$

Defining  $c_{jart}$  the values of  $c_j$  calculated for the artificial variables, those (non random) variables that give  $c_j$  such as:

$$|c_j| < \left| \max c_{jart} \right| \quad (49)$$

are considered not significant for the regression purpose and eliminated.

#### *Truncation-PLS*

Truncation-PLS can be based on different regression parameters; in this work it is based on loading weights, as in [76].

The starting point of the method is that, if a variable is uncorrelated to the response, loading weights are equally distributed random variables (not diverse from random normal noise).

Instead, if the **X**-variables are correlated to **Y**, loading weights are normally distributed as well but with non-zero mean. Hence, established a confidence interval around the median of the loading weights, it is possible to eliminate everything inside those intervals. Here the outlier detection has been conducted following the so-called *Lenth approach*, as suggested in [76].

#### *Forward selection*

The forward selection approach consists in starting with no variables in the model and then testing the inclusion of each variable. The process is repeated until the addition of a further variable does not improve the model. In particular, in this work, the inclusion of variables stops when the RMSECV does not decrease significantly with a further addition (the significance is checked by CVANOVA [77] with a confidence level of 95%).

#### *Jackknifing*

The jackknifing is a resampling procedure used for significance testing.

The uncertainty ( $s$ ) of a specific parameter is evaluated iteratively leaving out one observation at a time [78]. Then, uncertainty is estimated as:

$$s = \sqrt{\frac{\Omega-1}{\Omega}} \sum (\theta_i - \bar{\theta})^2 \quad (50)$$

Where  $\Omega$  is the number of models, and  $\theta_i$  and  $\bar{\theta}$  are the  $i$ -th estimate of the parameter and the mean of the parameter over the different models, respectively.

In this work, the uncertainty has been based on PLS-regression coefficients and calculated following the method proposed by Martens *et alia* in [79]:

$$\sum_m^M ((\mathbf{B} - \mathbf{B}_B) \zeta)^2 \quad (51)$$

Where  $\mathbf{B}$  and  $\mathbf{B}_B$  are the regression coefficient using all the objects and the regression coefficients using all the object except those left out in the  $m$ -th segment, respectively.

$\zeta$  is a scaling factor equal to:  $\sqrt{\frac{\Omega-1}{\Omega}}$ .

Regression coefficients of each variable are tested to be significantly different from zero. The variables related to those that result different (from zero), are selected.

## 4.2 Introduction of a variable selection step building multi-block models

As explained above, this work tries to cover, at least in part, the lack present in literature concerning the implementation of variable selection in multi-block model building. For the reasons explained above and in paper II, the different procedures proposed for combining Selectivity ratio, VIP and forward selection with MB-PLS and SO-PLS are the following:

- 1) MB-PLS combined with VIP
- 2) MB-PLS combined with SR
- 3) SO-PLS with pre-selected variables using VIP on each block
- 4) SO-PLS with pre-selected variables using SR on each block
- 5) SO-PLS combined with VIP
- 6) SO-PLS combined with SR
- 7) SO-PLS combined with forward selection

### **1-2) MB-PLS combined with VIP or SR**

In this procedures, MB-PLS models are (re)calculated after VIP or SR are used to reduce the number of variables of the predictor blocks. Following the MB-PLS procedure described in Paragraph 2.3.1, block-scaled predictor blocks are concatenated and then PLS is performed on the resulting matrix  $\mathbf{X}_{conc}$ . Variable selection (VIP or SR) is performed on this PLS model, obtaining the reduced matrix  $\mathbf{X}_{Red}$ . Subsequently, a new MB-PLS model is calculated using the reduced matrix  $\mathbf{X}_{Red}$ .

### **3-4) SO-PLS with pre-selected variables using VIP or SR on each block**

In these approaches the variable selection is performed on the individual blocks and then the SO-PLS model is built on the reduced blocks. Namely, the response is fitted to  $\mathbf{X}$  and  $\mathbf{Z}$  by two different PLS regressions. Then, VIP or SR are used to reduce the two predictor blocks, obtaining  $\mathbf{X}_{Red}$  and  $\mathbf{Z}_{Red}$ . Finally, the SO-PLS model is built using the reduced blocks. If wanted, it is also possible to select variable only in one block, and leave the other one unchanged.

### **5-6) SO-PLS combined with VIP or SR**

Due to the sequential nature of SO-PLS, variable selection can be integrated into the algorithm. In this case, after the first PLS is performed (step 1 in paragraph 2.3.3), variables are selected by VIP or SR and then the response is fitted to  $\mathbf{X}_{Red}$ . Subsequently,  $\mathbf{Z}$  is orthogonalized with respect to the scores of the previous PLS and used to predict the residual matrix from that regression (namely, the one involving  $\mathbf{X}_{Red}$ ). Also  $\mathbf{Z}_{Orth}$  is reduced by VIP or SR (obtaining  $\mathbf{Z}_{Orth,Red}$ ) and then the residual matrix is fitted to  $\mathbf{Z}_{Orth,Red}$ . Finally, the full predictive model is calculated combining the contributions from the regressions which involve the  $\mathbf{X}_{Red}$  and the  $\mathbf{Z}_{Orth,Red}$ . Also these two procedures could be applied to only one predictor block. If only the  $\mathbf{X}$  block is reduced (and  $\mathbf{Z}$  is left untouched), results obtained from procedures 3 and 5 and 4 and 6 coincide.

### **7) SO-PLS combined with forward selection**

SO-PLS has been combined with forward selection, in order to have regression models based on a reduced set of variables. The procedure can be applied following the steps below:

- 1) Any variable (from either  $\mathbf{X}$  or  $\mathbf{Z}$ ) is used to predict  $\mathbf{y}$  by PLS regression. The “best” predictor variable is selected looking at the RMSECVs obtained. It could be either an  $\mathbf{X}$ - or a  $\mathbf{Z}$ -variable.
- 2) Calling  $\Psi$  the total number of variables in both blocks ( $\Psi = J + K$ ),  $\mathbf{y}$  is fitted  $\Psi - 1$  times, using the variable selected in step 1 and any one of the other variables (one variable per time). This means that  $\mathbf{y}$  is predicted using only two predictor variables. If both variables come from the same block,  $\mathbf{y}$  is predicted by PLS regression, otherwise (one variable from  $\mathbf{X}$  and one from  $\mathbf{Z}$ ),  $\mathbf{y}$  is predicted by SO-PLS regression. The selected variable is the one which, together with the predictor selected in step 1, leads to the best prediction (lowest RMSECV) of  $\mathbf{y}$ .
- 3) PLS or SO-PLS models are built adding one variable per time to the two previously selected. The addition of all the variables (from both  $\mathbf{X}$  and  $\mathbf{Z}$ ) is tested. The selected variable is the one involved in the regression resulting in the lowest RMSECV. This step is repeated until any further addition improves predictions. The significance of the addition is checked by CVANOVA [77].

These procedures were tested in a simulation study, on a sensory data set and on a spectroscopic (Raman) data set. Performing variable selection, one could be particularly focused on having a reduced set of variables or pointing out the most meaningful variables in the system. From this study emerged that, independently of which one of these two is the aim of the variable selection, procedures based on SO-PLS appeared the most suitable in both cases. In particular, SO-PLS combined with forward selection is the best choice to obtain the most reduced sub-set of variables (still achieving acceptable predictive accuracy) for both the sensory and the spectroscopic data sets. When the focus is picking the most relevant variables (for prediction and interpretation), the suggested choice is different and more depending on the nature of the data set. In particular, VIP has demonstrated to be particularly efficient in getting rid of noise. This makes VIP the suitable choice in case of noisy data sets. On the other hand, selectivity ratio appears to be more effective in reducing/eliminating the systematic errors related to the presence of interferents. These characteristics may be used as a sort of guideline to indicate whether one or the other should be preferred. For instance, if one is interested in predicting the amount of all the constituents of a mixture based on the measured signal (e.g. spectroscopic or chromatographic), then VIP should be the most suitable approach. On the other hand, if the emphasis is placed on the prediction of a single

analyte, the others being considered as interferences, then selectivity ratio should be the most appropriate choice.

## ***Chapter 5: Conclusion***

The present study has tried to make a significant leap forward in the area of multi-block analysis. Crucial aspects that were not yet covered have been investigated and this has led to a number of important findings. Particular attention has been given to the use of multi-block methodologies in connection with food science. An extension of a pre-existing method to the classification field (SO-PLS-LDA) and a novel method for both regression and classification (SO-N-PLS/ SO-N-PLS-LDA) have been proposed. Practical applications of these to different fields of food science (e.g., chemical analysis, sensory assessments) have been reported. SO-(N)-PLS' performances have been inquired and they have been compared with well-known and widely applied methods. The novel approaches developed in the present thesis introduce some novelties (e.g., a graphical interpretation tool for classification results) and represent a solution for some issues (e.g., those related to the unfolding step required handling multi-way arrays); or at least they are a valid alternative to competing state of the art methodologies (Paper I, Paper III, Paper IV).

The study covers aspects of the multi-block field which have not been widely explored, i.e. the link between multi-block methods and three-way data structures (Paper III) and between multi-block methods and variables selection techniques (Paper II). Different procedures to perform variable selection have been studied. Some suggestions are given dependently of the nature of the data and of the final aim of the variable reduction.

The last part of the work has been focused more on interpretation of multi-block models rather than on predictions. Discussing this topic, different aspects about the interpretation of model parameters have been pointed out.

### **5.1 Main results**

The project covered different topics and a quite wide area of research. The present research aimed to develop some new tools for prediction and interpretation purposes in the multi-block field. Additionally, the thesis was finalized to give a comprehensive look to data fusion. Consequently, part of the work was focused on method development, and part opens up a discussion on the interpretation of the discussed models. Some theoretical and practical aspects have been discussed and investigated. Different targets has been achieved; some others still need to be studied more extensively.

Concerning the method-development part of the study, three methods (SO-PLS-LDA, SO-N-PLS and SO-N-PLS-LDA) have been proposed, tested and discussed. All of them are contributing to their state of the art under different aspects.

SO-PLS-LDA is a classification method for multi-block data sets obtained combining SO-PLS with LDA. From the prediction point of view, it is comparable to MB-PLS-LDA. Its importance belongs mainly to the benefits intrinsically present in the SO-PLS algorithm. It is therefore particularly relevant for the interpretation point of view.

SO-N-PLS is a multi-block regression method which can handle multi-way predictor arrays avoiding unfolding. This method has been developed to overcome issues related to the unfolding step. It has been tested only on three-way arrays, but, in principle, it can be applied to any multi-way array. It has demonstrated to have prediction ability comparable with (or higher than) other state of the art methods. Additionally, it allows obtaining some graphical interpretation tools that take into account the original structure of the data. The same characteristics are retained by SO-N-PLS-LDA. Consequently, these methods represent a good contribution to a field that had not been much explored so far.

Despite a great part of the work has been dedicated to the development of novel tools, attention has been given also to the interpretation of the multi-block models.

In Paper I, together with SO-PLS-LDA, a graphical representation of classification results based on cross-validated predicted values in the calculation of canonical variates has been presented. This fits well with the LDA philosophy and it is developed to ease the interpretation of classification models. Note that this kind of representation is not constrained to the use of SO-PLS-LDA as classification method. Consequently, it represents a general contribution to the classification field.

Another aspect that has been investigated in order to enhance the interpretation of multi-block models, is how to include variable selection in the data fusion context. Different procedures and considerations on how to proceed handling different data sets has been discussed. In particular, two different scenarios have been taken into account: a sensory data set and a spectroscopic (Raman) data set. It has been shown that, in both context, variable selection represent a reasonable tool to ease the interpretation of complex models such as the multi-block ones.

Finally, a discussion over the interpretation of some multi-block model parameters has been started. A preliminary investigation based on simulated data sets has been carried out; it has to be stressed that results still need to be investigated deeply and the discussion should be made wider and validated on real data sets. From the simulation study, some indications on the



interpretation of loadings have been pointed out. A criterion to estimate the number of interpretable components in a model has been proposed.

In conclusion, many of the planned goals have been achieved contributing to the data fusion context in different ways. In particular, the novel regression/classification methods could be easily used in the food industry to check the quality of foodstuff. Some MATLAB routines are already publicly available ([www.nofimamodeling.org](http://www.nofimamodeling.org)) and a reasonable future perspective would be to realized user friendly GUI to make them even more widely utilizable by workers of the sector.

## **5.2 Future Perspectives**

As mentioned in the previous paragraph, one obvious future perspective would be to make all the novel methodologies developed available in GUIs in order to make them easily usable in the industries. All the tools discussed in the thesis are thought for being applied in relation with the food industry (e.g., quality check, or predictions on consumer acceptance).

Nevertheless, the presented tools could be applied in several other domains. Multi-block regression and classification models are used in several field of science e.g. medical sciences, environmental studies, cosmology, or several others [80].

More conceptually speaking, interpretation of multi-block models definitely need further investigation. The achievements reached in Paper IV can be considered only indications, and need to be tested on real data sets and investigated/discussed deeply before they can be assumed to be generally valid and applicable.

Another topic to be further investigated is the complexity required by the methods. It would be interesting to carry out a simulation study varying the amount of noise in data and see whether and how the required complexity changes as a consequence.

Further perspectives cover also the algorithmic part of the study. In particular, it is natural to extend all the SO-PLS-based methods to the case of three (or more) block of predictors. Models based on several blocks should be investigated and probably additional tools to interpret these (more) complex data should be created.

Finally, a more challenging (but definitely worth investigating) perspective would be to study and develop a variable selection method suitable for SO-N-PLS (and multi-way regression methods in general). Firstly, the same procedures exposed in Paper II could be applied on unfolded three-way data to see if results vary from those obtained for SO-PLS. Then, model parameters could be studied to see if it is possible to find indices of the relevance of variables in the two (experimental/non-sample) modes in the prediction of the response. In this way the

selection of variables could be conducted on the original (multi-way) data (probably) granting the benefits due to the avoid of the unfolding step. It has to be stressed that the selection of variables on a three-way array is much more complex than on a two-way data matrix. In fact, to keep the multi-way data structure, the relevance of each variable in a mode should be tested in combination with all the predictors in the other; this may lead to the need of finding a compromise between parsimony and removal of potentially useful information.

## Chapter 6: Paper Summaries and References

### 6.1 Summary of Paper I

The focus of the paper is inspecting the possibility of using multi-block methods for classification purposes. In particular, the aim was to extend the SO-PLS regression method combining it with LDA. Consequently, the resulting method has been called SO-PLS-LDA. The novel method has been tested on a simulated data set and a real one. Both prediction and interpretation aspects are discussed. SO-PLS-LDA has been compared with other two classification methods: PLS-LDA and MB-PLS-LDA. The novel method show good results compared with the other multi-block approach. Both are definitely better than the individual block approach. Moreover, a graphical interpretation tool based on canonical variates is presented.

### 6.2 Summary of Paper II

The focus of this paper is to discuss and propose different procedures to perform variable selection in a multi-block context. In particular, the attention has been focused on two multi-block regression methods: Multi-Block Partial Least Squares (MB-PLS) and Sequential and Orthogonalized Partial Least Squares (SO-PLS). Firstly, seven variable selection methods were taken into account. A simulation study was conducted for regular PLS regression, selecting the variable selection methods to consider further in the multi-block context. Three methods were selected: Variance Importance in Projection (VIP) Selectivity Ratio (SR) and forward selection. Seven different procedures to combine MB-PLS and SO-PLS with these variable selection methods have been thoroughly examined. The benefits of performing variable selection can be better predictions and/or clearer interpretation. Both these aspects have been inspected. The different strategies have been tested out on simulation studies and on two different real data sets (one sensory and one spectroscopic data set).

### 6.3 Summary of Paper III

Due to the progress in modern instrumentations, it is becoming common to handle multi-way data. In order to make these data suitable for the classical data analysis, often they are unfolded. Unluckily, the unfolding procedure leads to some issues. Thus, a multi-way version of SO-PLS has been developed. The proposed method is called SO-N-PLS. It is an extension of SO-PLS where PLS is replaced by N-PLS. Obviously, this leads to other differences,

regarding both the algorithm and the structure of the parameters involved. In the paper, the method is exposed for three-way blocks, but it can be applied to any kind of multi-way array. SO-N-PLS has been tested in a simulation study and on two real data sets (in two cases in prediction and in one case in classification). Pros and cons of the method and its comparison with other multi-block methodologies (SO-PLS and MB-PLS on unfolded data sets) are described in the paper.

#### **6.4 Summary of Paper IV**

This paper is a discussion-oriented paper focused on interpretation of multi-block regression models. In particular, the discussion is restricted to SO-PLS, MB-PLS and PO-PLS methods and based on a simulation study. These methods are applied to solve regressions between blocks with different number of variables and different underlying components. Simulations are used to understand the reasonability of the interpretation of the loadings estimated from the different methods. On this regard, a specific approach to evaluate the “interpretability” of models is proposed. It has been called *explained variance criterion*.

## 6.5 References

- [1] M. Juran, A. Blanton, Juran's quality handbook, Fifth Edition McGraw-Hill, New York, NY, 1999
- [2] R. Bro, F. van den Berg, A. Thybo, C. M. Andersen, B. M. Jørgensen, H. Andersen, Multivariate data analysis as a tool in advanced quality monitoring in the food production chain, *Trends in Food Science & Technology*, 13 (2002) 235-244.
- [3] L. Munck, L. Nørgaard, S.B. Engelsen, R. Bro, C.A. Andersson, Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance, *Chemometr. Intell. Lab. Syst* 44, (1998), 31-60.
- [4] N. E. Tzouros, I. S. Arvanitoyannis Agricultural Produces: Synopsis of Employed Quality Control Methods for the Authentication of Foods and Application of Chemometrics for the Classification of Foods According to Their Variety or Geographical Origin, *Crit. Rev. Food Sci. Nutr.* 41 (2001) 287-319.
- [5] E. Borrás, J. Ferrer, R. Boque, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment- A review, *Anal. Chim. Acta* 891 (2015) 1-14.
- [6] M. Silvestri, L. Bertacchini, C. Durante, A. Marchetti, E. Salvatore, M. Cocchi, Application of data fusion techniques to direct geographical traceability indicators, *Anal. Chim. Acta* 769 (2013) 1– 9.
- [7] M. J. Dennis, Recent developments in food authentication, *Analyst*, 123 (1998) 151– 156.
- [8] I. E. Frank and B. R. Kowalski, Prediction of wine quality and geographic origin from chemical measurements by Partial Least-Squares regression modeling, *Anal. Chim. Acta*, 162 (1984) 241–251.
- [9] T. Skov, A.H. Honoré, H.M. Jensen, T. Næs, S.B. Engelsen, Chemometrics in foodomics: Handling data structures from multiple analytical platforms, *Trends Anal. Chem.* 60 (2014) 71-79.
- [10] R. Bro, PARAFAC. Tutorial and applications, *Chemometr. Intell. Lab. Syst* 38 (1997) 149-171.
- [11] P. M. Kroonenberg, J. de Leeuw, Principal components analysis of three-mode data by means of alternating least squares algorithms, *Psychometr.* 45 (1980) 69.

- [12] E. Acar, E. E. Papalexakis, G. Gurdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, R. Bro, Structure-Revealing Data Fusion, *BMC Bioinformatics*. 15 (2014) 239.
- [13] A.K. Smilde, J. Westerhuis, R. Boqué, Multiway multiblock component and covariates regression models, *J. Chemometr.* 14 (2000) 301-331.
- [14] T. Næs, O. Tomic, B. H. Mevik, H. Martens, Path modelling by sequential PLS regression, *J. Chemometr.* 25 (2011) 28–40.
- [15] J.A. Westerius, T. Kourti, J.F. MacGregor, Analysis of hierarchical PCA and PLS models, *J. Chemometr.* 12 (1998) 301–321.
- [16] G. Mazerolles, M. Hanafi, E. Dufour, D. Bertrand, E.M. Qannari, Common Components and Specific Weights Analysis: a chemometric method for dealing with complexity of food products, *Chemometr. Intell. Lab. Syst* 81 (2006) 41–49.
- [17] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, *J. Chemometr.* 25 (2011) 441–455.
- [18] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: Separating common and unique information in several data blocks, *Food Qual. Pref.* 24 (2012) 8–16.
- [19] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom, H. Wold, *Proc. Symp. On PLS Model Building: Theory and Application*, Frankfurt am Main, 1987; also Tech. rep., Department of Organic Chemistry, Umea University (1987).
- [20] M. Schouteden, K. Van Deun, S. Pattyn, and I. Van Mechelen. Sca with rotation to distinguish common and distinctive information in linked data. *Behav. Res. Methods* 45 (2013) 822-833.
- [21] E.F. Lock, K.A. Hoadley, J.S. Marron, and A.B. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl. Stat.* 7 (2013) 523-542.
- [22] M. Hanafi, A. Kohler, M. Qannari, Shedding new light on Hierarchical Principal Component Analysis, *J. Chemometr.* 24 (2010) 703–709.
- [23] El Ghaziri, V. Cariou, D.R. Rutledge, M. Qannari, Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of  $(K + 1)$  datasets, *J. Chemometr.* 30 (2016) 420-429.
- [24] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, *J. Chemometr.* 10 (1996) 463–482.

- [25] I. E. Frank, J. Feikema, N. Constantine and B. R. Kowalski, *J. Chem. Info. Comput. Sci.* 24 (1984) 20–24.
- [26] I. E. Frank and B. R. Kowalski, Prediction of wine quality and geographic origin from chemical measurements by Partial Least-Squares regression modeling, *Anal. Chim. Acta*, 162 (1984) 241–251.
- [27] L. E. Wangen and B. R. Kowalski, A multiblock Partial Least Squares algorithm for investigating complex chemical systems *J. Chemometrics* 3 (1988) 3–20.
- [28] J.A. Westerhuis, A.K. Smilde, Deflation in multiblock PLS, *J. Chemometr.* 15 (2001) 485-493.
- [29] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, *J. Chemometr.* 15 (2001) Pages 715-742.
- [30] Jørgensen K, Mevik B-H, Næs T. Combining designed experiments with several blocks of spectroscopic data. *Chemometr. Intell. Lab. Syst* 88 (2007) 154–166.
- [31] I Måge, B. H Mevik, T. Næs, Regression models with process variables and parallel blocks of raw material measurements, *J. Chemometr.* 22 (2008) 443–456.
- [32] S. Bougeard, M. Qannari, N. Rose, Multiblock redundancy analysis: Interpretation tools and application in epidemiology, *J. Chemometr.* 25 (2011) 467-475.
- [33] S. Bougeard, M. Qannari, C. Lupo, M. Hanafi, From Multiblock Partial Least Squares to Multiblock Redundancy Analysis. A Continuum Approach, *Informatica* 22 (2011) 11-26.
- [34] H. Hotelling. Relations between two sets of variates, *Biometrika*, 28 (1936) 321–377.
- [35] U. Indahl, A twist to partial least squares regression, *J. Chemometr.* 19 (2005) 32–44.
- [36] M. Seasholtz, B. Kowalski, Qualitative information from multivariate calibration models, *Appl. Spectrosc.* 44(1990) 1337–1348.
- [37] K. Kjeldahl, R. Bro, Some common misunderstandings in chemometrics, *J. Chemometrics* 24 (2010) 558–564.
- [38] R. Bro, Multi-way calibration. Multilinear PLS, *J. Chemometr.* 10 (1997) 47-61.
- [39] A.K. Smilde, R. Bro, P. Geladi, Multi-way analysis: Application in the chemical sciences, John Wiley and Sons, New York, NY, 2004.
- [40] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, New York, 1989.
- [41] S. De Jong, Short communication regression coefficients in multilinear PLS, *J. Chemometr.* 12 (1998) 77-81.

- [42] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, R. Nescatelli, F. Marini, Classification and class-modeling. In: F. Marini (Ed.), *Chemometrics in Food Chemistry*, Elsevier, Oxford, UK, 2013, pp.171-233.
- [43] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd Ed., Wiley, New York, NY, 2000.
- [44] C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, Four levels of pattern recognition, *Anal. Chim. Acta*, 103 (1978) 429-443.
- [45] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, *Chemometr. Intell. Lab. Syst.* 93 (2008) 132–148.
- [46] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, *Trends Anal. Chem.* 35 (2012) 74-86.
- [47] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* 7 (1936) 179–188.
- [48] W. Krzanowski, *Principles of multivariate analysis*, 2nd Ed., Oxford University Press, Oxford, UK, 2000.
- [49] G. McLachlan, *Discriminant analysis and statistical pattern recognition*, Wiley, New York, NY, 2004.
- [50] H. Hotelling, The most predictable criterion. *J. Educ. Psych.* 26 (1935) 139-142.
- [51] A. Lubin, Linear and non-linear discriminating functions. *Br. J. Psych. Stat. Sect.* 3 (1950) 90-103.
- [52] R.B. Darlington, S.L. Weinberg, H.J. Walberg, Canonical Variate Analysis and Related Techniques, *Rev. Educ. Res.* 43 (1973) 433-454.
- [53] L. Nørgaard, R. Bro, F. Westad, S.B. Engelsen, A modification of canonical variates analysis to handle highly collinear multivariate data. *J. Chemometr.* 20 (2006) 425-435.
- [54] M. Barker, W. Rayens, Partial least squares for discrimination. *J. Chemometr.* 17 (2003) 166.
- [55] H. Nocairi, E.M. Qannari, E. Vigneau, D. Bertrand, Discrimination on latent components with respect to patterns Application to multicollinear data, *Comput. Stat. Data Anal.* 48 (2005) 139.
- [56] U.G. Indahl, H. Martens, T. Naes, From dummy regression to prior probabilities in PLS-DA. *J. Chemometr.* 21 (2007) 529–536.



- [57] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, Mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, *Chemometr. Intell. Lab. Syst.* 137 (2014) 181–189.
- [58] R.S. Halinski, L.S. Feldt, The selection of variables in multiple regression analysis, *J. Educ. Meas.* 7 (1970) 151–157.
- [59] M.L Thompson, Selection of variables in multiple regression: Part I. A review and evaluation, *Int. Stat. Rev.* 46 (1978) 1–19.
- [60] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Reducing over-optimism in variable selection by cross-model validation, *Chemometr. Intell. Lab. Syst.* 84 (2006) 69–74.
- [61] S. Osborne, R. Künnemeyer, R. Jordan, Method of wavelength selection for partial least squares, *Analyst*, 122 (1997) 1531–1537.
- [62] H. Swierenga, P.J. de Groot, A.P. de Weijer, M.W.J. Derksen, L.M.C. Buydens, Improvement of PLS model transferability by robust wavelength selection, *Chemometr. Intell. Lab. Syst.* 41 (1998) 237–248.
- [63] J.-P. Gauchi, P. Chagnon, Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, *Chemometr. Intell. Lab. Syst.* 58 (2001) 171–193.
- [64] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, *J. Mach. Learn. Res.* 3 (2003) 1157–1182.
- [65] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, *Chemometr. Intell. Lab. Syst.* 118 (2012) 62–69.
- [66] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, *J. Chemom.* 24 (2010) 728–737.
- [67] L. Nørgaard, A. Saudland, J. Wagner, J. Nielsen, L. Munck, S. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, *Appl. Spectrosc.* 54 (2000) 413–419.
- [68] F. Savorani, M.A. Rasmussen, Å. Rinnan, S.B. Engelsen, Interval-Based Chemometric Methods in NMR Foodomics. In: F. Marini (Ed.), *Chemometrics in Food Chemistry*, Elsevier, Oxford, UK, 2013, pp.171–233.
- [69] T. Rajalahti, R. Arnenberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, *Chemom. Intell. Lab. Syst.* 95 (2009) 35–48.

- [70] O.M. Kvalheim, Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots, *J. Chemometr.* 24 (2010) 496–504.
- [71] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, *Anal. Chem.* (2009) 2581–2590.
- [72] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures. In: H. Kubinyi (Ed.) *3D QSAR in drug design: theory, methods and applications*, ESCOM Science Publishers, Leiden, The Netherlands, 1993, pp.523–550.
- [73] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, *Chemom. Intell. Lab. Syst.* 129 (2013) 76–86.
- [74] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC), *Chemom. Intell. Lab. Syst.* 138 (2014) 153–160.
- [75] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, *Anal. Chem.* 68 (1996) 3851–3858.
- [76] K.H. Liland, M. Høy, H. Martens, S. Sæbø, Distribution based truncation for variable selection in subspace methods for multivariate regression, *Chemom. Intell. Lab. Syst.* 122 (2013) 103–111.
- [77] U. Indahl, T. Næs, Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling, *J. Chemom.* 12 (1998) 261–278.
- [78] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982 (ISBN 0–89871–179-7).
- [79] H. Martens, M. Martens, Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression, *Food Qual. Prefer.* 11 (2000) 5–16.
- [80] D. Lahat, T. Adali, C. Jutten, Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects *Proc. IEEE* 103 (2015) 144.

## **Part II:**

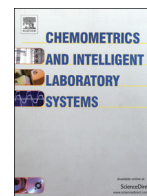
# **Publications**

## **Paper I**

**Title:** Combining SO-PLS and linear discriminant analysis for multi-block classification

**Authors:** A. Biancolillo, I. Måge, T. Næs

Published in Chemometrics and Intelligent Laboratory Systems, 141 (2015) 58–67.



# Combining SO-PLS and linear discriminant analysis for multi-block classification

Alessandra Biancolillo<sup>a,b,\*</sup>, Ingrid Måge<sup>a</sup>, Tormod Næs<sup>a,b</sup>

<sup>a</sup> Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway

<sup>b</sup> Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

## ARTICLE INFO

### Article history:

Received 28 August 2014

Received in revised form 1 December 2014

Accepted 4 December 2014

Available online 16 December 2014

### Keywords:

SO-PLS

Multiblock

Linear discriminant analysis

Regression

Classification

## ABSTRACT

The aim of the present work is to extend the Sequentially Orthogonalized-Partial Least Squares (SO-PLS) regression method, usually used for continuous output, to situations where classification is the main purpose. For this reason SO-PLS discriminant analysis will be compared with other commonly used techniques such as Partial Least Squares-Discriminant Analysis (PLS-DA) and Multiblock-Partial Least Squares Discriminant Analysis (MB-PLS-DA). In particular we will focus on how multiblock strategies can give better discrimination than by analyzing the individual blocks. We will also show that SO-PLS discriminant analysis yields some valuable interpretation tools that give additional insight into the data. We will introduce some new ways to represent the information, taking into account both interpretation and predictive aspects.

© 2014 Elsevier B.V. All rights reserved.

## 1. Introduction

In recent years, the advancement of new analytical technologies has generated an increased need for interpreting large and complex data sets and also the relationship between them. Often, variables can be separated into conceptually meaningful blocks of data which can represent for instance measurements taken by different instruments or at different time points in a process. This kind of multi-block data can be found in various fields, such as industrial processes monitoring, consumer and sensory science, in the -omics area, in microbiology and in medical protocols [1,2]. It is therefore crucial to develop and implement methods which can process this large amount of data and that allow for combining multiple blocks of data from different experimental conditions. Several useful methods have been proposed for the purpose, for instance Multiblock-PLS, Multiblock-Principal Component Analysis (MB-PCA), SO-PLS, Parallel Orthogonalized Partial Least Squares (PO-PLS) and OnPLS [3–6], but the area is still new and there are many unsolved problems. These range from issues related to experimental design and variable selection to how to handle non-linearities, interactions and classification. In the present paper the focus will be

on extending regression oriented multi-block methodology to the area of multi-block classification.

The aim of this work is to extend the SO-PLS regression method, until now only used for continuous output, to situations where classification is the main purpose. In order to do that, SO-PLS is combined with Linear Discriminant Analysis (LDA). How this can be done in practice is discussed and demonstrated in two examples. An important aspect is to focus on how multiblock strategies can give better discrimination than analyzing the individual blocks separately. Main focus will be on classification ability, but the importance of interpretation will also be highlighted. Interpretation tools of more general interest for classification will also be proposed. The classification ability of the SO-PLS-LDA will also be compared with other standard techniques such as PLS-LDA and MB-PLS-LDA. Two data sets will be used for illustration; a data set established for distinguishing between wines from different countries and a data set based on simulations.

## 2. Theory and methods

We will start with a short description of discriminant analysis based on one data matrix ( $\mathbf{X}$ ), where  $\mathbf{X}$  consists of  $N$  objects (rows) and  $p$  variables (columns). The  $N$  objects come from  $J$  different classes, with  $N_j$  samples in each class. We then proceed with describing discriminant analysis based on multiple data blocks, where an additional data matrix  $\mathbf{Z}$  (with dimensions  $N \times r$ ) is also available for describing the differences

\* Corresponding author at: Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway. Tel.: +47 64 97 01 15.

E-mail address: [alessandra.biancolillo@nofima.no](mailto:alessandra.biancolillo@nofima.no) (A. Biancolillo).

between classes. Finally, we will propose some novel tools for statistical testing of contributions from  $\mathbf{X}$  and  $\mathbf{Z}$ , and some graphical interpretation tools for the multiblock models.

### 2.1. Discriminant analysis

Classification is the process of assigning objects to a set of different classes or categories. Methods that require a training set where the categories are known in advance are called “supervised methods” or “discriminant analysis”, which will be the focus in this paper.

#### 2.1.1. Linear discriminant analysis (LDA)

One of the oldest discriminant analysis methods is called *linear discriminant analysis* (LDA) and was proposed by Fisher [7] in 1936. This method is based on the assumption that the probability distribution within each class follows Gaussian (normal) distribution. In addition to the normality assumption, the LDA assumes that a priori probabilities  $\pi_j$  for each of the  $J$  classes are defined. These probabilities can for instance be estimated by the training set as  $N_j/N$  or set equal for each class as  $\frac{1}{J}$ . In this paper will use the second approach. Following Bayes rule, each sample is assigned to the group with the highest posterior probability. With the above assumptions, this implies that each sample is assigned to the class  $j$  that gives the smallest value of  $C_j$ :

$$C_j = (\mathbf{x}_i - \mu_j)^T \Sigma^{-1} (\mathbf{x}_i - \mu_j) + \log|\Sigma| - 2 \log(\pi_j) \quad (1)$$

where  $\mu_j$  s are the class means and  $\Sigma$  is the variance/covariance matrix common to all the classes (the so-called *pooled* variance/covariance matrix). Note that with equal prior probabilities for each class this criterion reduces to a Mahalanobis distance only.

The mean and covariance matrices need to be estimated from the data. The means are usually calculated as the group means  $\bar{\mathbf{x}}_j$ . For the common covariance matrix  $\Sigma$ , the following estimate is usually used:

$$\mathbf{S} = \sum_{j=1}^J \frac{(N_j - 1) \mathbf{S}_j}{(N - J)} \quad (2)$$

where  $\mathbf{S}_j$  is the empirical variance/covariance matrix for class  $j$ .

The main limitation of LDA is that it requires a well-conditioned covariance matrix. This means that the method cannot be used when the number of variables exceeds the number of samples, or when the variables themselves are highly correlated. To overcome this limitation, methods that use latent variables have been proposed. One of these methods, that will be discussed in the next session, is the *partial least squares–discriminant analysis* (PLS-DA).

**2.1.1.1. Canonical variates.** It can be demonstrated that, in a two category case, there is only one specific direction that gives the maximum separation between the classes, i.e. which maximizes the ratio of the distances between the means of classes and the variances within each class. This idea can be generalized to several classes, ending up with linear combinations of the data that separate the classes as much as possible. The  $\mathbf{S}_b$  between-group variance/covariance matrix is defined as:

$$\mathbf{S}_b = \sum_{j=1}^J (\bar{\mathbf{x}}_j - \bar{\mathbf{x}}) (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T \quad (3)$$

where  $\bar{\mathbf{x}}_j$  and  $\bar{\mathbf{x}}$  represent the mean vector of class  $j$  and of the whole dataset, respectively.

In the general multi-category case, canonical variates ( $\mathbf{w}_i$ ) can be calculated by solving the generalized eigenvalue problem:

$$\mathbf{S}^{-1} \mathbf{S}_b \mathbf{w}_i = \lambda_i \mathbf{w}_i \quad (4)$$

As for the LDA formula, the canonical variates are sensitive to overfitting and one needs to be careful when interpreting them. In Section 2.3.3 we will therefore propose an alternative procedure based on the predictions from the cross-validation.

#### 2.1.2. PLS-DA and PLS-LDA

PLS-DA removes the drawbacks arising from an ill-conditioned covariance matrix since it is based on the transformation of original variables into a smaller number of latent, orthogonal variables. In PLS-DA, the  $\mathbf{Y}$  matrix used as response is a particular matrix consisting of 0's and 1's (called *Dummy*) containing the information about the class membership [8–11].

The variability present in the blocks  $\mathbf{X}$  and  $\mathbf{Y}$  is described by two sets of latent variables: the scores  $\mathbf{T}$  and  $\mathbf{U}$ , which are chosen in the way to maximize their covariance. The model structure for PLS regression can be written as [11–13]:

$$\mathbf{X} = \mathbf{T} \mathbf{P}^T + \mathbf{E}_X \quad (5)$$

$$\mathbf{Y} = \mathbf{U} \mathbf{Q}^T + \mathbf{E}_Y \quad (6)$$

$$\mathbf{U} = \mathbf{T} \mathbf{K} \quad (7)$$

where  $\mathbf{T}$  and  $\mathbf{U}$  are the scores of the blocks  $\mathbf{X}$  and  $\mathbf{Y}$ ,  $\mathbf{P}$  and  $\mathbf{Q}$  are the respective loadings,  $\mathbf{E}_X$  and  $\mathbf{E}_Y$  are the residual matrices that represent the unexplained variance, and  $\mathbf{K}$  is the diagonal matrix of the regression coefficients of the linear relationship between the scores; Eq. (7) is the one called *Inner relation*. Recombining these expressions it is possible to obtain the matrix of the regression coefficients  $\mathbf{B}$  that is necessary for the prediction of the vector  $\hat{\mathbf{Y}}_{\text{New}}$ .

$$\hat{\mathbf{Y}}_{\text{New}} = \mathbf{X}_{\text{New}} \mathbf{B} \quad (8)$$

Classification for the standard PLS-DA is then usually done by assigning the unknown sample to the class corresponding to the column of  $\hat{\mathbf{Y}}_{\text{New}}$  that has the largest value [8,9].

Another possible way of using the information in the PLS-DA model was suggested by Indahl et al. [8,14]. Instead of using the highest value of the predicted  $\mathbf{Y}$ , one can apply LDA on the PLS scores. In other words, the PLS with dummy  $\mathbf{Y}$  is used as a data compression before the use of the standard LDA. This method will here be called PLS-LDA. As shown in [8], this method allows extracting relevant and stable components for classification. Moreover, it represents a way to overcome problems concerning the dimensionality. Therefore, this will be the approach to be pursued below. Note that it can be shown that doing LDA on the scores from the PLS regression is identical to doing LDA on predicted values from PLS regression [8].

### 2.2. Discriminant analysis with multiple data sets

When considering multi-block methods, we will focus on the two-block case in this paper, but the methodologies can easily be expanded to any number of blocks. We assume the following linear model structure:

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{Z} \mathbf{C} + \mathbf{E} \quad (9)$$

where  $\mathbf{Y}$  is the dummy matrix representing class belonging,  $\mathbf{X}$  and  $\mathbf{Z}$  are the two descriptive data blocks, and  $\mathbf{E}$  contains residuals.

**Table 1**  
Contingency table.

	Method 1 correct	Method 1 wrong	Row total
Method 2 correct	A	B	A + B
Method 2 wrong	C	D	C + D
Column total	A + C	B + D	N

### 2.2.1. Sequentially orthogonalized PLS (SO-PLS) discriminant analysis

The SO-PLS regression [14] extracts information sequentially from each data block, which means that the chosen order of the blocks can influence the end result. The SO-PLS algorithm starts by fitting  $\mathbf{Y}$  to  $\mathbf{X}$  by PLS regression. In this step  $\mathbf{X}$ -scores ( $\mathbf{T}_x$ ),  $\mathbf{X}$ - and  $\mathbf{Y}$ -loadings ( $\mathbf{P}_x$  and  $\mathbf{Q}_x$ , respectively) and the matrix of the residual ( $\mathbf{E} = \mathbf{Y} - \mathbf{T}_x \mathbf{Q}_x^T$ ) are calculated.

After that, the  $\mathbf{Z}$  block is orthogonalised with respect to the scores of the previous PLS:

$$\mathbf{Z}_{\text{orth}} = \mathbf{Z} - \mathbf{T}_x (\mathbf{T}_x^T \mathbf{T}_x)^{-1} \mathbf{T}_x^T \mathbf{Z} \quad (10)$$

and then  $\mathbf{Z}_{\text{orth}}$  is fitted to the  $\mathbf{Y}$ -residuals ( $\mathbf{E}$ ). In this way, it is possible to extract further information from  $\mathbf{Z}$  that explains the remaining variance in  $\mathbf{Y}$ . Since  $\mathbf{Z}$  can be decomposed into a contribution projected onto the PLS scores of  $\mathbf{X}$  (i.e.  $\mathbf{T}_x$ ) and a contribution orthogonal to the PLS scores of  $\mathbf{X}$ , the column spaces spanned by  $\mathbf{T}_x$ ,  $\mathbf{Z}$  and  $\mathbf{T}_x$ ,  $\mathbf{Z}_{\text{orth}}$  are the same (see also reference [3]). This means that from a prediction point of view the orthogonalization does not represent any loss of information. In the following and last step, since  $\mathbf{T}_x$  and  $\mathbf{T}_z^{\text{orth}}$  are orthogonal due to the orthogonalization in the second step, the full predictive model is obtained by adding the two contributions, i.e.:

$$\hat{\mathbf{Y}} = \mathbf{T}_x \mathbf{Q}_x^T + \mathbf{T}_z^{\text{orth}} (\mathbf{Q}_z^{\text{orth}})^T \quad (11)$$

where the  $\mathbf{Q}$ 's are the regression coefficients. The full predictive model can also be written as function of the original measures, and it can be computed as explained in [15].

It should, however, be emphasized that the row spaces of  $\mathbf{T}_x$ ,  $\mathbf{Z}$  and  $\mathbf{T}_x$ ,  $\mathbf{Z}_{\text{orth}}$  are not the same. In other words, the  $\mathbf{Z}_{\text{orth}}$  is not in the space spanned by  $\mathbf{Z}$ . This may possibly be seen as a drawback from an interpretation point of view (for  $\mathbf{Z}$ ) with the SO-PLS approach as it

stands now, but the consequences have not yet been explored. This is clearly a place where improvement may be possible and more research is needed. For this paper, only tools related to the original SO-PLS proposal are used.

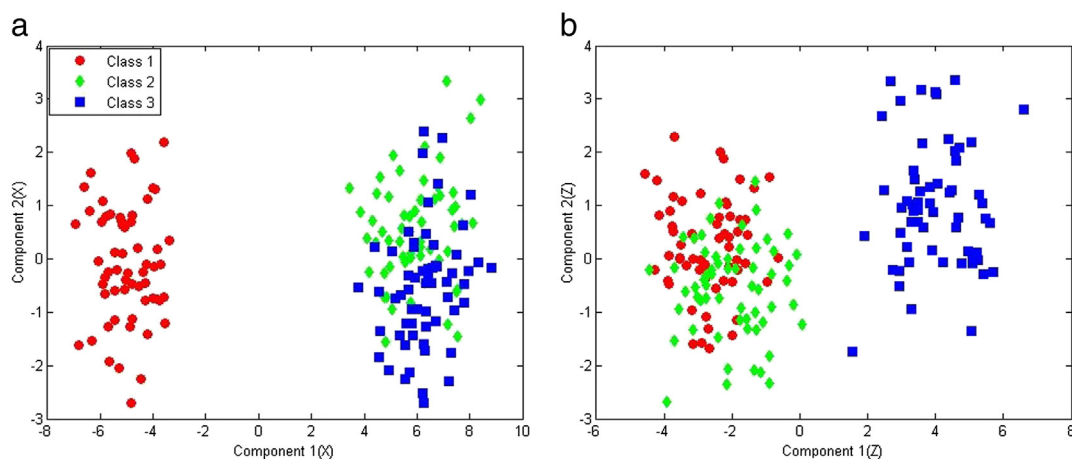
The natural tools for interpretation of the first PLS model (for  $\mathbf{X}$ ) are the scores and loadings for the first PLS model. For the second model it is also natural to use the PLS scores for  $\mathbf{Z}_{\text{orth}}$ , representing the additional information obtained by incorporating  $\mathbf{Z}$  in the regression model, but in this case it is better (see also [15]) to look at the projection of  $\mathbf{Z}$  onto the PLS scores than considering the PLS loadings (for  $\mathbf{Z}_{\text{orth}}$ ) themselves. The reason for this is that in this way one obtains a more direct interpretation of how the real  $\mathbf{Z}$  relates to the extra variability from  $\mathbf{Z}$  that is incorporated in the model.

The number of latent variables is decided independently for each block, usually by cross-validation. There are two strategies for selecting the number of components: sequential or global [15]. The latter is the one used in this work. In the global approach the best possible combination is determined evaluating a so-called Måge plot [15]. This is a graph which shows all the possible combinations of latent variables (LVs) reporting the RMSECVs as a function of the total number of components. For practical use of the methods one should as always test the method with the selected number of components on an independent data set.

An important advantage of SO-PLS is that it is invariant to block scaling and it provides interpretation tools for investigating both the contribution of  $\mathbf{X}$ , the additional information of block  $\mathbf{Z}$ , as well as their joint influence on  $\mathbf{Y}$  [3]. The  $\mathbf{X}$ -block is typically interpreted by the PLS model itself while the additional contribution of  $\mathbf{Z}$  may be interpreted by projecting the true  $\mathbf{Z}$  onto the PLS scores of  $\mathbf{Z}_{\text{orth}}$ , showing how the true  $\mathbf{Z}$  related to the additional information extracted after  $\mathbf{X}$  has been modeled. It can also explicitly handle situations with different numbers of underlying components in each of the blocks, which may potentially be an advantage for understanding better the individual dimensionalities of the blocks. Adding more blocks than two can easily be done by repeating orthogonalization with respect to the scores of the previous PLS regression, and fitting the orthogonalized block with the preceding residual matrices.

### 2.2.2. Multiblock PLS (MB-PLS) discriminant analysis

The standard MB-PLS (Westerhuis et al. [16]) approach consists in concatenating the input blocks and then performing PLS regression on the resultant matrix. This means that the units used for the different blocks will have an effect on the solution. To overcome this possible drawback, blocks are usually block-scaled. When MB-PLS is used for



**Fig. 1.** Distribution of samples in the first two component of both  $\mathbf{X}$ - and  $\mathbf{Z}$ -block. a) In the  $\mathbf{X}$ -block Class 1 (red dots) is separated from the other two along the first component while all the classes are overlapped on the second one. b) In the  $\mathbf{Z}$ -block Class 3 (blue squares) is separated from the other two along the first component while all the classes are overlapped on the second one and even on the third (not shown). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

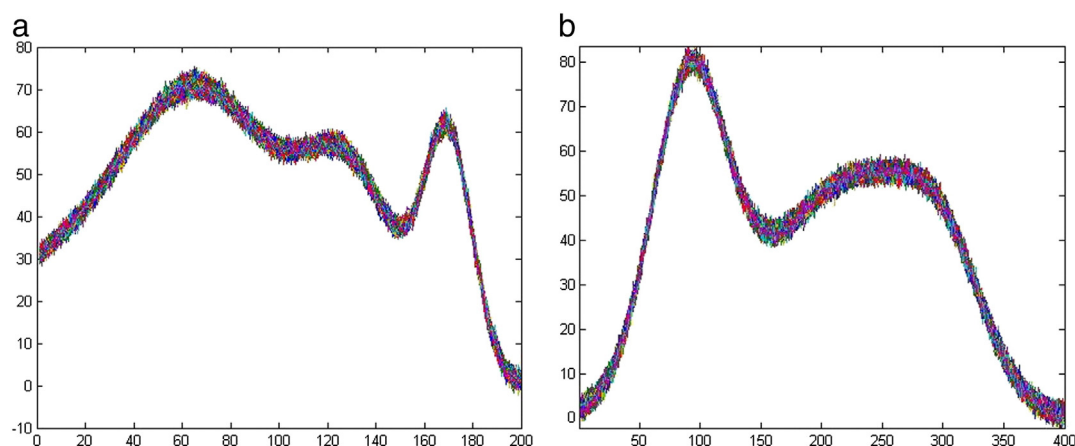


Fig. 2. Simulated dataset: a) Block **X** training and **X** test b) Block **Z** training and **Z** test.

classification, LDA is applied on the super-scores, and not the block-scores. The method will be called MB-PLS-LDA.

In this work, in order to build the MB-PLS-LDA model, both blocks are divided by their respective Euclidean norms and then concatenated. LDA is then applied to the super-scores. The general experience is that MB-PLS regression is easy to calculate and implement, and that it gives good results [16–19].

### 2.3. Interpretation tools

In addition to the canonical variates plots discussed above and regular plotting of scores and loadings for the PLS models obtained, we propose the following tools for interpretation.

#### 2.3.1. Comparing correct and wrong classifications

The most obvious way of comparing methods is to set up a table as indicated in Table 1. In that case both correct and wrong classifications are counted for two different methods, for instance a one-block and a multi-block method.

A more detailed study considering all blocks and their classification error can also be undertaken as will be discussed when presenting Table 6 below. The idea is most interesting when comparing a full SO-PLS-LDA model with models based on **X** and **Z** separately. For all samples that are correctly classified by SO-PLS-LDA, one can count how many are correctly and erroneously classified for **X** and **Z**

separately. The same can be done for the samples which are erroneously classified by the full SO-PLS-LDA model.

#### 2.3.2. Statistical testing of improvements incorporating **Z**

In order to evaluate the actual usefulness of the addition of the **Z**-block different statistical tests will be used here. Note that these tests are always based on results from cross-validation and they are as such analogous to the use of the CV-ANOVA test of additional contributions for the continuous **Y** case described in Næs et al. [3]. It should be mentioned that this test could also be used here, but it is more natural to use the more directly interpretable tables below.

In this paper we will focus on the Mc Nemar Test [20] (with Yates's continuity correction [21]). This is a statistical test usually applied to  $2 \times 2$  contingency tables and it is used to compare the proportions of paired data (see Table 1). It can be used to test for statistical significance of the null hypothesis of equality of classification error of two methods (here called method 1 and method 2). In our case, the two methods will be classification methods based on two different types of input, **X** and **X, Z**.

The hypothesis that is relevant to test here is that the two methods have the same classification error. The test statistic  $\chi^2$  is calculated as stated in the following equation:

$$\chi^2 = \frac{(B-C)^2}{B+C} \quad (12)$$

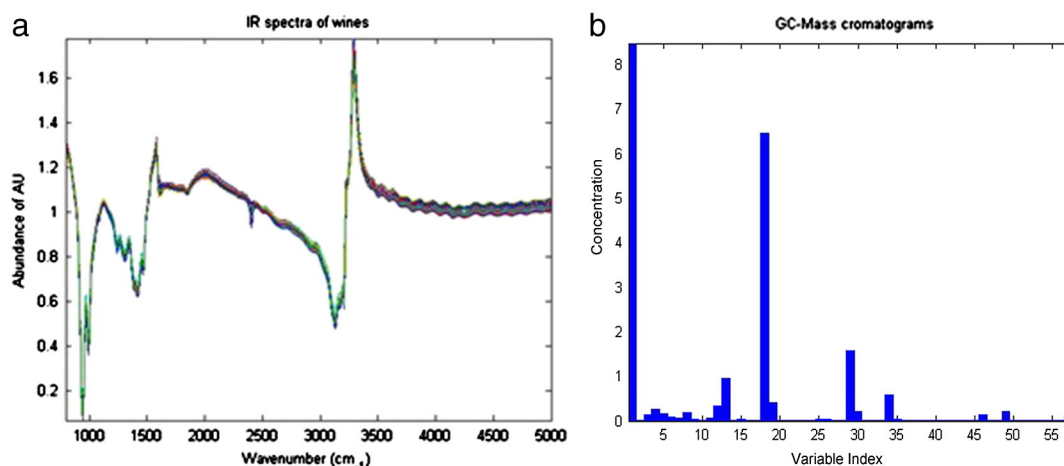
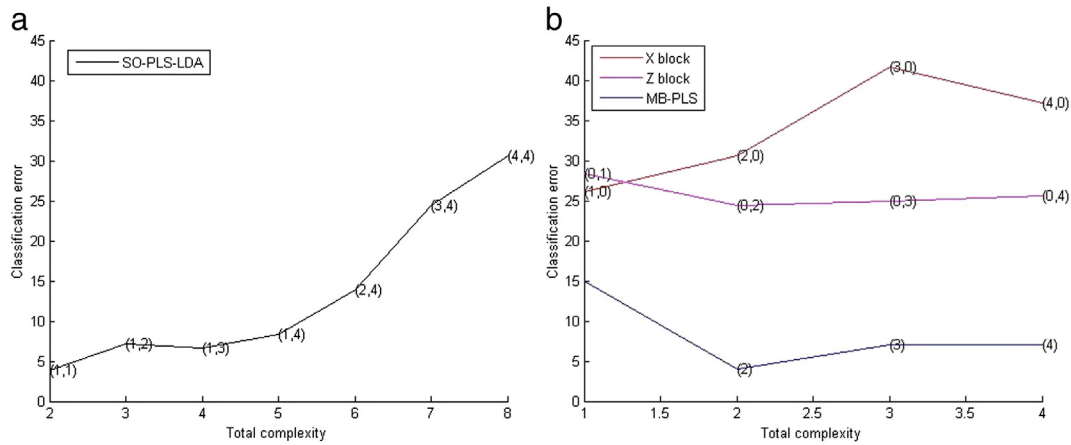


Fig. 3. a) IR spectra of wines. b) Bar representation of the Aroma compounds chromatogram.





**Fig. 4.** Måge Plot: a) Values of classification error when different numbers of components are used in SO-PLS-LDA. b) Values of classification error when different numbers of components are used in MB-PLS-LDA and when PLS is applied on only **X**- or only **Z**-block.

where  $B$  and  $C$  are as defined in the table above. When values of frequency distribution are small, it is recommended to use the Yates's continuity correction; so Eq. (12) becomes:

$$\chi^2 = \frac{(|B-C|-0.5)^2}{(B+C)} \quad (13)$$

On the base of the  $\chi^2$ 's value the null hypothesis is accepted or rejected.

McNemars' test is simple to use, but suffers from the fact that it is less suitable for small data sets. For this reason, also Fisher's exact test can be performed [22]. This is a statistical test of significance used in the analysis of categorical data when chi-square test can't be used.

### 2.3.3. Canonical variates based on cross-validation

Above we stated that canonical variates calculated the normal way can give an overoptimistic view on classification ability. In the following we will propose an alternative based on cross-validation Y-values. The method can be used for any of the approaches discussed above.

The procedure goes as follows:

Let us call  $\hat{Y}$  the matrix of cross-validated-predicted Y-values, whose dimensions are  $N \times J$ . This can be considered as a partitioned matrix constituted by  $J$  submatrices of cross-validated-predicted Y-values, one for each of the categories in the training set.

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_J \end{bmatrix} \quad (14)$$

**Table 2**  
Classification error for PLS-LDA performed on the only X- and Z-block (Test sets).

X-block				
PLS-LDA				
LVs	Class	Predicted class 1	Predicted class 2	Predicted class 3
1	1	60	0	0
	2	0	28	32
	3	0	26	34
Z-block				
PLS-LDA				
1	1	44	10	6
	2	4	56	0
	3	5	0	55

For each of these submatrices  $\hat{Y}_j$ , the mean prediction vector (centroid) can be defined as  $\bar{y}_j^T$  and calculated by taking the average of each column in  $\hat{Y}_j$ . Accordingly, the  $S$  covariance is calculated as in Eq. (2) where  $S_j$  is:

$$S_j = (\hat{Y}_j - 1\bar{y}_j^T)^T (\hat{Y}_j - 1\bar{y}_j^T) \quad (15)$$

While  $S_b$  is calculated as:

$$S_b = \sum_{j=1}^J (\bar{y}_j - \bar{y})(\bar{y}_j - \bar{y})^T \quad (16)$$

At the end, canonical variates are obtained solving Eq. (4).

Note that the canonical variates can be calculated on the predicted values since these are linear functions of the scores. In order to obtain full rank one column has to be discarded from the  $\hat{Y}$  matrix.

In this way, the plot will be based on objectively predicted values found without the influence from the true group membership and are therefore safer to look at than the standard canonical variates. This type of canonical variates can also be obtained in a test set by using predicted values.

### 2.4. Selection of PLS components

The number of components in PLS regression is usually based on the cross-validated root mean squared prediction error (RMSECV). For all the methods where PLS regression is combined with LDA, the optimal number of components can be selected based on either RMSECV (reflecting the dummy Y-matrix) or on the percentage of correct classifications (also cross-validated) from LDA. In principle, the latter is the one to be recommended since it better fits with the objective of classification. However, when the number of samples in each category is low, we have observed that the correct classification rates may become unstable and overfitted. This is due to the fact that the LDA corresponds to a crisp classification (either 0 or 1), while the predicted dummy matrix is analogous to a fuzzy classification (continuous values). For small data sets, we therefore recommend basing the estimation of the number of components on the RMSECV, and then use this when applying the LDA. An alternative which could also be of interest here is the methods proposed by Westerhuis et al. [23], but this is not pursued here.

In SO-PLS the number of components in each PLS can be defined using two different approaches: the *global optimization* and the *sequential optimization* [15]. The approach pursued in this work is the global one. In order to verify if the two approaches give different results,

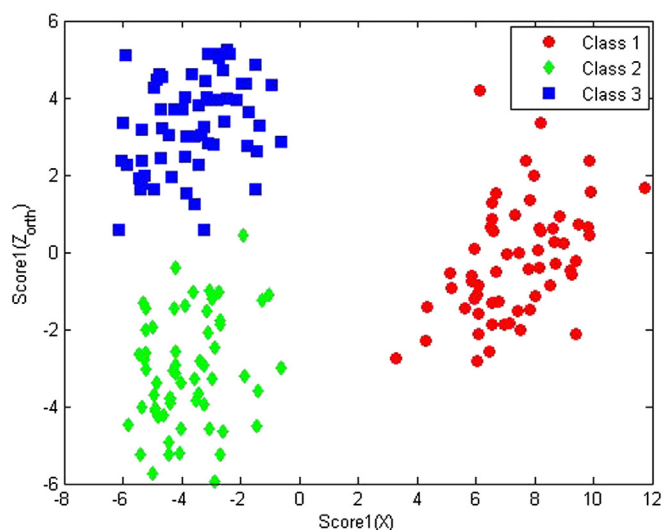


Fig. 5. SO-PLS Model. Samples projected in the space of the first  $X$ - and  $Z_{\text{orth}}$ -scores.

the sequential optimization has also been calculated for both datasets. It leads to the selection of the same number of components as the global approach.

### 3. Data sets

#### 3.1. Simulated dataset

The simulated data set consists of three groups and two spectroscopic blocks. Two  $X$ -blocks and two  $Z$ -blocks (one for the training and one for the test set) with spectra-like loadings have been generated.

The  $X$ -block was designed as a two-component system in which separation between class 1 and classes 2 and 3 occurs along the first component. The three classes are overlapped along the second axis. Three components were used to simulate the  $Z$ -block, which was designed to have perfect separation between class 3 and the other two categories along the first component. The three classes overlap along the other two components. A further visualization of the distribution of samples along components 1 and 2 is shown in Fig. 1.

The relative variances of the  $X$ -components are 96.3% (for the first one) and 3.7% (for the second one), while the variances of the  $Z$ -components are 79.6%, 9.8% and 10.6% for each, respectively.

For both the  $X$  and  $Z$  blocks, spectra-like loadings were built as orthogonalized combinations of Gaussians functions. These Gaussians (six for the  $X$ -block and ten for the  $Z$ -block) with different means and

random variances are used. After the building of loadings, they are orthogonalized and they are divided by their Euclidean norm. Therefore, each block is built of not necessarily orthogonal scores but the loadings of each block are orthonormal. At last 5% (of the average of the reconstructed spectral signal) Gaussian noise was added to the matrices after reconstruction from the components space. The blocks  $X$  are  $180 \times 200$  matrices while the blocks  $Z$  are  $180 \times 400$ . For both training and test sets, samples were divided into three groups of equal size. Simulated spectra are reported in Fig. 2.

#### 3.2. Wine data

38 samples of red wines, produced from the same grape (100% Cabernet Sauvignon), harvested in different geographical areas, have been collected from local supermarkets in the area of Copenhagen, Denmark. Wines are produced in Australia, Chile, and South Africa. Of the 38 samples, 12 are from Australia, 15 from Chile and 13 from South Africa. The wines are characterized by two different data blocks: *FT-IR spectra* (842 variables spanning the wavelength region 5011–929  $\text{cm}^{-1}$ ) and *Aroma compounds* (estimated by integration of 57 selected GC-MS peaks). The objective was then to see if the measured data can be used to discriminate between geographical origins.

This is a subset of a larger data set, and detailed description of the samples and laboratory measurements can be found in [24].

In Fig. 3, the spectra and the aroma compounds are plotted. In Fig. 3a are shown merely the IR spectra of all the samples. In Fig. 3b is shown the bar plot of the averages of the concentrations of the various aroma compounds in the different samples. The aroma compounds were scaled to unit variance prior to analysis, as their variation ranges were considerably different.

#### 3.3. Data analysis

All data analysis was performed using MATLAB (r2012b, The Mathworks, Natick, MA), using in-house routines for PLS-LDA, SO-PLS-LDA and MB-PLS-LDA. The MATLAB routines are available for download at [www.nofimamodeling.org](http://www.nofimamodeling.org).

### 4. Results

Both data sets were first analyzed by PLS-LDA using only  $X$  or  $Z$  as input blocks. Then, SO-PLS-LDA was applied using both  $X$ - and  $Z$ -blocks. In the end, MB-PLS-LDA was performed for comparison. For all these models, the same validation criterion and the same way of choosing the number of latent variables were used. In LDA, the a priori probability was set to  $1/J$  for all classes.

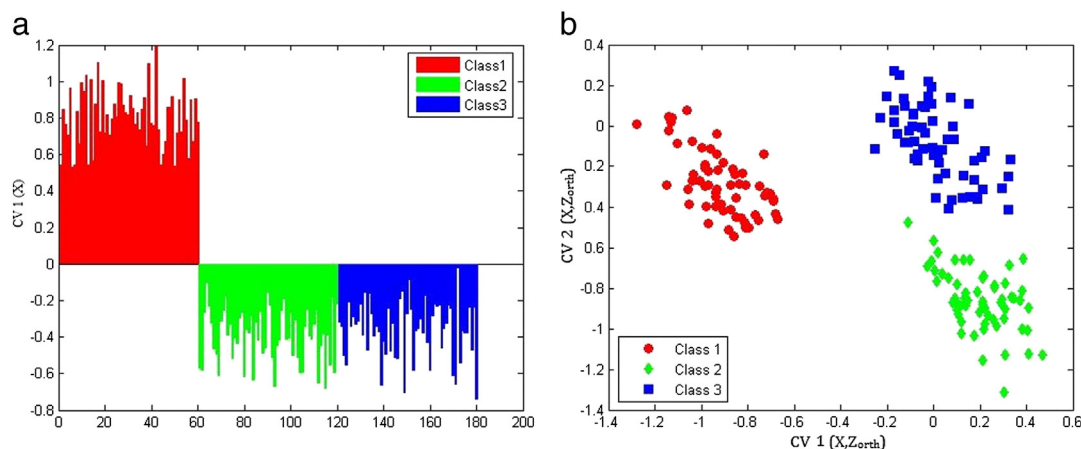


Fig. 6. Crossvalidated scores of the training samples onto the relevant canonical variates: a) PLS-LDA on  $X$ -block (Training set); b) SO-PLS-LDA on both  $X$ - and  $Z$ -blocks (Training set).

**Table 3**

Contingency table on simulated data. Method 1 stands for SO-PLS-LDA on both X- and Z-block and Method 2 stands for PLS-LDA for X-block only.

	Method 1 correct	Method 1 wrong	Row total
Method 2 correct	122	0	122
Method 2 wrong	58	0	58
Column total	180	0	180

Since both datasets proposed have more than two classes, the Y block contains variables without an intrinsic low-dimensional structure. Therefore, the adaptation proposed by Barker and Rayens [10] has been considered. Overall classification errors obtained applying SO-PLS-LDA with or without this correction are the same (see also [8]) and this method is therefore not pursued further here.

#### 4.1. Results on dataset 1: simulated data

##### 4.1.1. Optimal complexity

This data set is large enough to allow model optimization based on LDA classification instead of RMSECV as will be used for the data set below. The model was optimized using cross-validation with seven cancellation groups. The Måge plot based on classification errors in Fig. 4 shows that the lowest error for SO-PLS-LDA is given by 1 latent variable for the first PLS and 1 for the second one. The 0 solutions do not seem relevant for the choice of SO-PLS components so they are not shown in the plot. For PLS-LDA, 1LV is used both for the X- and Z-blocks and two LVs were selected for MB-PLS-LDA.

##### 4.1.2. PLS-LDA classification using one block only

In order to assess the classification error using only one block, a regular PLS-LDA on the X- and Z-block was performed and then the classification error is calculated on predictions based on the test set. For the X-block the overall classification error is 32% (58 misclassified samples) while for the Z-block the global error is 14% (25 misclassified samples). Test set results related to each specific class are reported in Table 2. It is clear from the table that the X-block is able to separate class 1 from the other two, while the Z-block is able to separate class 3 from the other two. Both these results are as expected according to how the data are generated.

##### 4.1.3. SO-PLS-LDA and MB-PLS-LDA

The two multiblock methods allow making much better classifications than the previous ones. In particular, SO-PLS-LDA reaches the 100% of correct classification. MB-PLS-LDA misclassifies one sample

**Table 4**

Classification error of PLS-LDA performed on the only Aroma block and on the only IR block.

X-block: Aroma (A)				
PLS-LDA				
LVs	Class	Predicted "Australia"	Predicted "Chile"	Predicted "South Africa"
4	1	9	1	2
	2	1	12	2
	3	0	3	8
Z-block: IR (I)				
PLS-LDA				
6	1	8	2	3
	2	0	14	1
	3	3	1	6

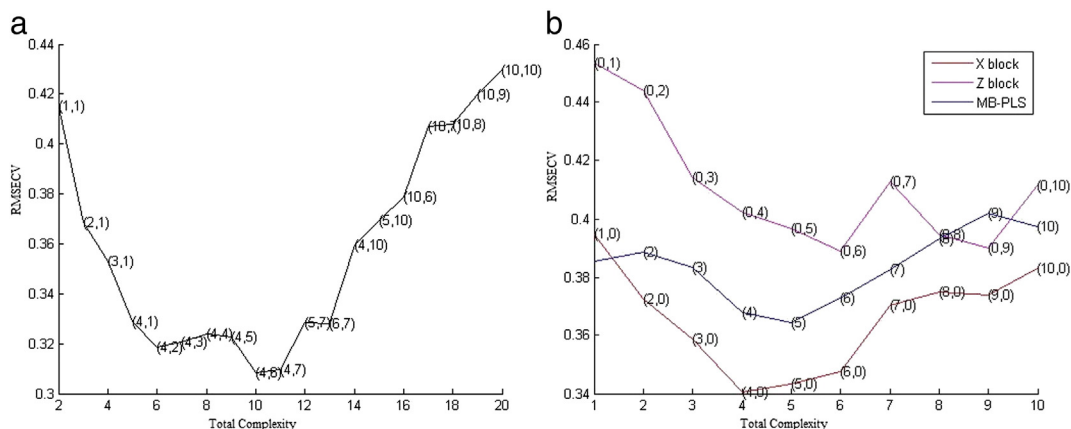
belonging to class 2 assigning it to the class 3 and one sample from class 3 assigning it to the class 2. Therefore, the global error for this method is 1%. In other words, the methods are very comparable for this data set. As can be seen, when putting together the X- and Z-block in a single model, the classification becomes perfect (or close to perfect for the MB-PLS-LDA).

##### 4.1.4. Visual inspection of classification results

In Fig. 5 are reported samples in the space of the first X- and Z<sub>orth</sub>-Scores for the SO-PLS-LDA. The figure clearly shows that only one component from each of the groups is enough for separating the groups perfectly. This corresponds very well to how the data were generated; one component in X for separating class 1 from the other two and one component in Z for separating mainly class 3 from the other two. As can also be seen, there is some overlap between class 1 and class 3 along the second axis, but in the two dimensional space, the three classes are completely separated.

As already proposed, a natural way for interpreting and visualizing SO-PLS-LDA (and other) models is to use the canonical variates approach (based on predicted values from cross-validation or prediction testing on independent data).

Fig. 6a shows projections along the first canonical variates when PLS-LDA is applied on only one block (X-block). Note that only one canonical variate can be calculated here due to the fact that only one component is used in the model. It is clear that class 1 can be separated by the other two on this direction but it is not possible to discriminate class 2 from class 3.



**Fig. 7.** a) Values of RMSECV when different numbers of components are used in SO-PLS-LDA. b) Values of RMSECV when different numbers of components are used in MB-PLS-LDA and when PLS is applied on only X- or only Z-block. The figure is based on a model built using the peak areas of the Aroma compounds as X-block and the IR spectra as Z-block. On the SO-PLS-LDA curve the first number is the number of latent variables chosen for the first PLS and the second one is the number of latent variables chosen for the second PLS. The solutions with 0 components in either Z or X are omitted from Fig. 7a since they are presented in Fig. 7b and since they are also not relevant for the selection of components.

**Table 5**

Comparison between SO-PLS-LDA and MB-PLS-LDA on wines.

X-block: Aroma (A); Z-block: IR (I); (model AIY)				
LVs	Class	Predicted "Australia"	Predicted "Chile"	Predicted "South Africa"
<i>SO-PLS-LDA</i>				
4,2	1	9	1	2
	2	1	14	0
	3	1	2	8
<i>MB-PLS-LDA</i>				
4	1	9	1	2
	2	1	14	0
	3	1	1	9

Fig. 6b on the other hand shows crossvalidated scores of the samples onto the two canonical variates when SO-PLS-LDA is applied on both **X**- and **Z**-block. The plot shown is based on predicted values from cross-validation. As can be seen, the three classes are perfectly separated. The first canonical variate separates between class 1 and the other two, while the second one separates class 2 from other two. The same plot has been made using predicted values for the test set samples and it is not reported here as it is very similar to the plot in Fig. 6b.

#### 4.1.5. Testing improvements incorporating Z

Although the multi-block approach is obviously much better than using only one block, we incorporate the above mentioned contingency table and the Mc Nemar test. Table 3 shows the actual number of correct and incorrect classifications. Method 1 means SO-PLS-LDA and Method 2 means the PLS-LDA used for **X** only.

The calculated Mc Nemar test statistics is in this case equal to  $\chi^2 = 57.004$ . The tabulated value of  $\chi^2$  for one degree of freedom at 95% of confidence is 3.84. This would mean that the two can be considered extremely statistically different, as expected.

The other table tool that is mentioned above does not provide much additional insight when the differences are as obvious as here. So we leave further visual inspection to the next section.

## 4.2. Results on dataset 2: wines

### 4.2.1. Optimal complexity

The data set was not considered large enough to use the LDA classification for selecting number of components, so the component selection was based on RMSECV as described in Section 2.4.

Fig. 7 shows the RMSECV as a function of numbers of components for SO-PLS-LDA, MB-PLS-LDA and PLS-LDA.

Concerning SO-PLS-LDA, the plot shows all the possible combinations of latent variables under a fixed maximum value. (In this case, the number of latent variables was limited to ten). Even in this case, it doesn't seem relevant to show the RMSECV when 0 components are selected for one of the two blocks in SO-PLS-LDA. The lowest value of RMSECV is obtained by taking four latent variables for the first PLS and six for the second one, but as can be seen, four and two components give almost the same results. We therefore decided to use four and two components in further comparisons.

For MB-PLS-LDA, four LVs were chosen. Concerning PLS-LDA on **X**- and **Z**-block the numbers of LVs used are set to four and six respectively.

### 4.2.2. PLS-LDA classification using one block only

In order to assess the classification error using only one block, a regular PLS-LDA on the Aroma block was performed. The overall classification error is 24% (9 misclassified samples). Classification results related to each specific class are reported in Table 4.

**Table 6**

Global view of results: comparison between LDA applied on only **X**- or **Z**-block after PLS-LDA and SO-PLS-LDA. The table shows how many samples correctly classified by SO-PLS-LDA (full model) are correctly/incorrectly classified by PLS-LDA and therefore, how many samples misclassified by SO-PLS-LDA are correctly/incorrectly classified by PLS-LDA. (PLS-LDA has been applied on only **X**-, **Z**- and **Z<sub>orth</sub>**-block).

	Full model correct: 31	Full model wrong: 7
X-only correct	29	0
X-only wrong	2	7
Z-only correct	23	5
Z-only wrong	8	2
Z <sub>orth</sub> -only correct	10	6
Z <sub>orth</sub> -only wrong	21	1

PLS-LDA was applied on the IR-block also. The model gives a global classification error of 26% (10 misclassified samples).

From Table 4 we see that the Aroma block is slightly better at discriminating class 1 and class 3, while the IR block is better at discriminating class 2.

### 4.2.3. SO-PLS-LDA and MB-PLS-LDA

As stated in [14] the order of the blocks in SO-PLS is not so important for prediction purposes, although it may be important for interpretation. In order to check if this statement is reasonable also in this case, the two different SO-PLS models with the opposite order of the blocks have been created. Even if the number of latent variables selected in the two models is different, both give the same number of misclassified samples. Hereafter, we will focus on the model using aroma as **X** and IR as **Z**.

The SO-PLS-LDA and MB-PLS-LDA results are reported in Table 5. Concerning SO-PLS-LDA, only seven samples are now misclassified which corresponds to a global classification error of 18%. MB-PLS-LDA instead misclassifies six samples, reaching a global classification error of 16%. As can be seen, results are somewhat better when using both blocks **X** and **Z** than when using only one of them. Again, the two multi-block methods are comparable. The two multi-block methods give similar results. More details can be found in Table 6. For such a small data set there is, however, such a difference is not significant.

It is interesting to note that even though Fig. 7 gives a lower RMSECV for **X** only than for MB-PLS, the actual classification results are better for the multi-block approach.

### 4.2.4. Testing improvements incorporating Z

The Mc Nemar test returns a value:  $\chi^2 = 1.125$ . This corresponds to a *p*-value of 0.239. The tabulated value of  $\chi^2$  for one degree of freedom at 95% of confidence is 3.84. As can be seen even though the differences are quite clear from the canonical variates plot, the classification improvement is not significant at 5% level. Fisher's exact test was also calculated. The *p*-value obtained from the test supports the same conclusion. A typical procedure for obtaining a more firm assessment of significance or not would be to increase the data set, but this is outside the scope of the paper.

### 4.2.5. Further analysis of tables

For the same reason exposed above, even in this case, it could be useful to have a table with a more detailed view of the results. These are reported in Table 6. Note that this is essentially a series of tables of the same type as Table 3.

From the table it is possible to see that results obtained from the Aroma block (indicated as **X**-only in Table 6) are similar to those obtained by SO-PLS-LDA. Indeed, it shows that of the thirty one samples correctly classified by SO-PLS-LDA on **X**- and **Z**-block, twenty nine are rightly classified even by PLS-LDA on **X**. But as can also be seen, two of the samples that were correctly classified by the full model were erroneously classified by using **X** only. The remaining seven samples are erroneously classified by both methods. It should be stressed that, even if the number



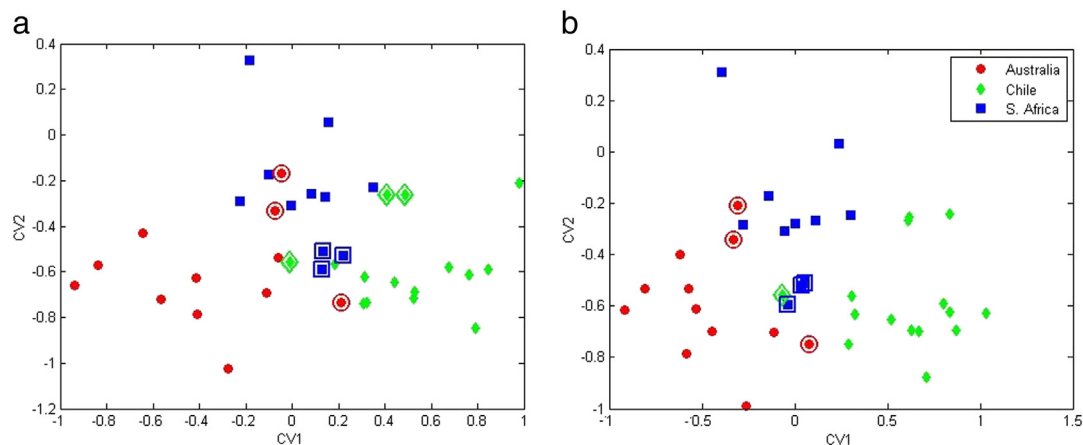


Fig. 8. Prediction in the space on canonical variates using only Aroma block (a) or Aroma and IR blocks (b). Samples marked by circles are the misclassified ones.

of incorrectly classified samples are the same, they are not always assigned to the same (wrong) class by the two models. Concerning the IR block ( $Z$ -only in Table 6), it is possible to observe that PLS-LDA on this block gives results more different from SO-PLS-LDA. For instance eight of the thirty one samples that were correctly classified by SO-PLS-LDA are erroneously classified by  $Z$  alone (hence, as reported in the fourth row of Table 6, eight of the thirty one samples properly classified by SO-PLS-LDA are misclassified by PLS-LDA on  $Z$ ). One can also see the somewhat interesting phenomenon that among the seven erroneously classified samples by SO-PLS-LDA, five of them were correctly classified by PLS-LDA on  $Z$ . This result also points to the possibility of combining the  $Z$ -blocks and  $X$ -block in an even better way, but this is not considered further here. Afterwards, in the last two rows of Table 6, even the results obtained when PLS-LDA is applied on  $Z_{orth}$  are reported. In this case, only ten of the thirty one samples correctly classified by SO-PLS-LDA are properly classified by PLS-LDA (and so twenty one of these thirty one are misclassified by PLS-LDA on  $Z_{orth}$ ). In conclusion, it is clear also here that the Aroma block dominates the model more than the IR block.

#### 4.2.6. Visual inspection of classification results

In Fig. 8 we present samples in the space of the canonical variates, when these are extracted by using the predicted values obtained by cross validation both for  $X$  and for  $X, Z$  used together.

From the plot it is clear that classes are better separated in SO-PLS-DA. In Fig. 8a samples assigned to different classes are overlapped. While in 8b overlapping between samples is almost completely avoided. Samples marked by a circle are the misclassified ones.

**4.2.6.1. Investigating sequential PLS models.** In order to get more deep insight in the modeling results, one can use scores and loadings plot from the  $X$  and  $Z$  model separately. Doing a full interpretation of the blocks is beyond the scope of the present paper, so here we just add two plots for illustrating the potential.

In Fig. 9a the scores plot for the  $Z_{orth}$  block is shown. There is a not clear grouping tendency along the two components, it is quite weak; but it can be investigated for interpretation purposes. In Fig. 9b is shown the loading plot. In this particular case, instead of discussing interpretation of the PLS model for  $Z_{orth}$ , which is not in the row space spanned by  $Z$ , we consider the projection of  $Z$  itself onto the scores of this PLS model. The procedure is simple, fits well with the philosophy of the method and it gives us a direct way of seeing how the original  $Z$  relates to the extra information obtained after  $X$  has been fitted (see also reference [3]). From this it is possible to identify the chemical profile of the components. In fact, the first two components clearly have an IR spectra-like shape and the main variables of the spectra are quickly identifiable. The most evident are the ones related to the

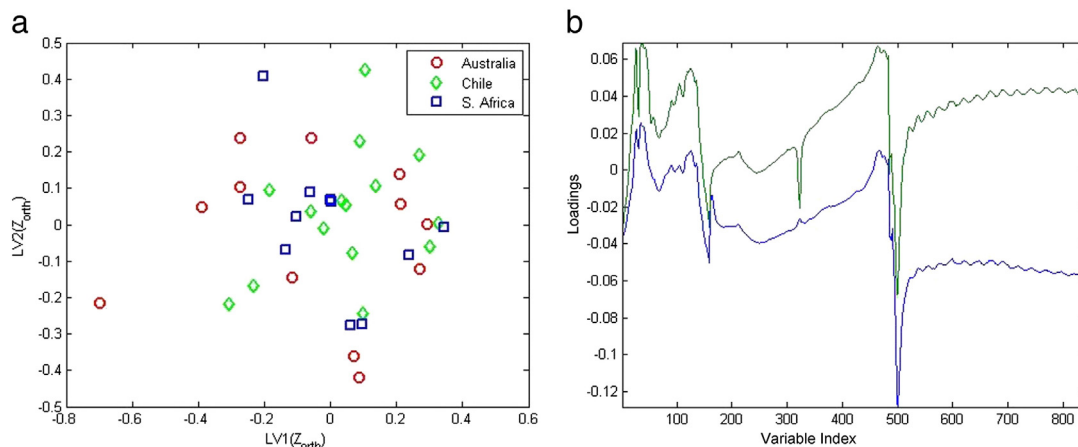


Fig. 9. a) Scores plot on the orthogonalized IR-block; b) loading plot of the IR-block.

stretching of the O–H and N–H bounds, or the absorptions given by the ester bond C–O.

## 5. Conclusions

In this paper, a novel classification method for multi-block data configurations was proposed as the combination of SO-PLS and LDA. Its characteristics and outcomes are compared to other existing approaches either involving a single data matrix or more than one block.

In general, it was found that, at least when the blocks have complementary information, multi-block methods give better results as compared to methods that use one block at a time. In the present work, the use of the multi-block approaches on one simulated and one real dataset resulted in better classification ability, as compared to the cases where PLS-LDA was used on the individual blocks. In particular, for both datasets, SO-PLS-LDA gives comparable accuracy to MB-PLS-LDA (based on combining regular multi-block PLS regression with LDA).

Even though the prediction ability of the two multi-block methods was comparable for the data sets tested here, the proposed SO-PLS-LDA possesses a number of properties with a potential benefit for prediction and interpretation, due to the orthogonalization. In particular, the SO-PLS-LDA does not need any type of block scaling since it is invariant to the relative scale of the blocks. Another advantage is that it can be used to compute and assess the number of components separately for the different blocks, meaning that it can easily handle for instance combinations of design variables (full rank blocks) and multicollinear blocks. This may have advantages both for prediction ability and interpretation of the different blocks. A power of the method is also its ability of directly assessing and visualizing separately the contributions of each block. We refer to reference [3] for a more general discussion of SO-PLS based interpretation. The advantage of the standard MB-PLS approach is, however, its simplicity and easy access to software. There are also a larger number of successful applications of the method, in particular for regression. A deeper study of the advantages and possible disadvantages of the two methods in light of these aspects with focus on both regression and classification, is needed.

A number of interpretation tools of general interest have been proposed and illustrated. In particular, different tools based on interpretation and statistical handling of data tables containing the number of correct/wrong classifications have been proposed. A more general tool based on using cross-validated predicted values in the calculation of canonical variates has been proposed. The advantage of this approach is that it does not base the canonical variates on overoptimistic results.

## Conflict of interest

The authors declare that they don't have any conflict of interest.

## Acknowledgments

We would like to thank FFL and the Research Council of Norway for financial support (Project number: 225096).

## References

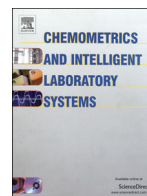
- [1] R. Bro, F. van den Berg, A. Thybo, C.M. Andersen, B.M. Jørgensen, H. Andersen, Multivariate data analysis as a tool in advanced quality monitoring in the food production chain, *Trends Food Sci. Technol.* 13 (2002) 235–244.
- [2] S. Hassani, H. Martens, E.M. Qannari, M. Hanafi, G.I. Borge, A. Kohl, Analysis of -omics data: graphical interpretation- and validation tools in multi-block methods, *Chemometr. Intell. Lab. Syst.* 104 (2010) 140–153.
- [3] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, *Chemometr. Intell. Lab. Syst.* 124 (2013) 32–42.
- [4] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, *Food Qual. Prefer.* 24 (2012) 8–16.
- [5] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, *J. Chemometr.* 25 (2011) 441–455.
- [6] E. Acar, E.E. Papalexakis, G. Gürdeniz, M.A. Rasmussen, A.J. Lawaetz, M. Nilsson, R. Bro, Structure-revealing data fusion, *BMC Bioinforma.* 15 (2014) 239.
- [7] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* 7 (1936) 179–188.
- [8] U.G. Indahl, H. Martens, T. Næs, From dummy regression to prior probabilities in PLS-DA, *J. Chemometr.* 21 (2007) 529–536.
- [9] H. Nocairi, E.M. Qannari, E. Vigneau, D. Bertrand, Discrimination on latent components with respect to patterns. Application to multicollinear data, *Comput. Stat. Data Anal.* 48 (2004) 139–147.
- [10] M. Barker, W. Rayens, Partial least squares for discrimination, *J. Chemometr.* 17 (2003) 166–173.
- [11] L. Ståle, S. Wold, Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study, *J. Chemometr.* 1 (1987) 185–196.
- [12] H. Martens, T. Næs, *Multivariate Calibration*, John Wiley & Sons New York, NY, 1991.
- [13] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS methods, in: A. Ruhe, B. Kågström (Eds.), *Matrix pencils: proceedings of a conference held at Pite Havsbad, Sweden, March 22–24, 1982*, Springer Verlag, Heidelberg, Germany, 1983, pp. 286–293.
- [14] U.G. Indahl, N.S. Sahni, B. Kirkhus, T. Næs, Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise, *Chemometr. Intell. Lab. Syst.* 49 (1999) 19–31.
- [15] T. Næs, O. Tomic, B.-H. Mevik, H. Martens, Path modelling by sequential PLS regression, *J. Chemometr.* 25 (2011) 28–40.
- [16] J.A. Westerius, T. Kourti, J.F. MacGregor, Analysis of hierarchical PCA and PLS models, *J. Chemometr.* 12 (1998) 301–321.
- [17] J. Westerhuis, P. Coenegracht, Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares, *J. Chemometr.* 11 (1997) 379–392.
- [18] J.F. MacGregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, Process monitoring and diagnosis by multiblock PLS methods, *Proc. Syst. Eng.* 40 (1994) 826–838.
- [19] S. Wold, S. Hellberg, T. Lundstedt, M. Sjöström, H. Wold, PLS modelling with latent variables in two or more dimensions, *Proc. Symp. on PLS Model Building: Theory and application*, Frankfurt, 1987.
- [20] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentage, *Psychometrika* 12 (1947) 153–157.
- [21] F. Yates, Contingency table involving small numbers and the  $\chi^2$  test, *J. R. Stat. Soc. Suppl.* 1 (1934) 217–235.
- [22] R.A. Fisher, On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P, *J. R. Stat. Soc.* 85 (1922) 87–94.
- [23] J.A. Westerhuis, E.J.J. van Velzen, H.C.J. Hoefsloot, A. Smilde, Discriminant  $Q^2$  ( $DQ^2$ ) for improved discrimination in PLS-DA models, *Metabolomics* 4 (2008) 293–296.
- [24] T. Skov, D. Balabio, R. Bro, Multiblock variance partitioning. A new approach for comparing variation in multiple data blocks, *Anal. Chim. Acta.* 615 (2008) 18–29.

## **Paper II**

**Title:** Variable selection in multi-block regression

**Authors:** A. Biancolillo, K. Hovde Liland, I. Måge, T. Næs, R. Bro

Published in Chemometrics and Intelligent Laboratory Systems, 156 (2016) 89–101.



## Tutorial Article

## Variable selection in multi-block regression

Alessandra Biancolillo<sup>a,b,\*</sup>, Kristian Hovde Liland<sup>a,c</sup>, Ingrid Måge<sup>a</sup>, Tormod Næs<sup>a,b</sup>, Rasmus Bro<sup>b</sup><sup>a</sup> Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway<sup>b</sup> Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark<sup>c</sup> Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N-1432 Ås, Norway

## ARTICLE INFO

## Article history:

Received 1 October 2015

Received in revised form 19 April 2016

Accepted 23 May 2016

Available online 27 May 2016

## Keywords:

Variable selection

Multi-block

SO-PLS

MB-PLS

Raman

Sensory

## ABSTRACT

The focus of the present paper is to propose and discuss different procedures for performing variable selection in a multi-block regression context. In particular, the focus is on two multi-block regression methods: Multi-Block Partial Least Squares (MB-PLS) and Sequential and Orthogonalized Partial Least Squares (SO-PLS) regression. A small simulation study for regular PLS regression was conducted in order to select the most promising methods to investigate further in the multi-block context. The combinations of three variable selection methods with MB-PLS and SO-PLS are examined in detail. These methods are Variable Importance in Projection (VIP) Selectivity Ratio (SR) and forward selection. In this paper we focus on both prediction ability and interpretation. The different approaches are tested on three types of data: one sensory data set, one spectroscopic (Raman) data set and a number of simulated multi-block data sets.

© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

With the advancement of technology, data collected in many fields of science are getting more informative, but at the same time also more complex. For example, analytical measurements can now typically be obtained with different instruments, in different places and at different times of a production process [1]. In consumer and sensory science, it is common that several data sets represent aspects that need to be considered together in order to obtain the information wanted [2]. Even in medical protocols, data can be represented by blocks of independent variables [3] that need to be considered together. Different multi-block methods have been proposed, e.g. Multiblock PCA, generalized Procrustes analysis, Multi-Block-PLS (MB-PLS), Sequential and Orthogonalized Partial Least Squares (SO-PLS), Parallel Orthogonalized Partial Least Squares (PO-PLS), OnPLS and others [4–9]. Multi-block analysis is still a young field and several problems and challenges are unsolved. One of these is variable selection for the purpose of improved interpretation and prediction in regression models.

Variable selection in regression can lead to a number of advantages. For instance, removing noisy or irrelevant variables may result in improved predictions and a reduction of the model complexity. Feature selection can also ease interpretation. From a practical point of view, selecting variables can make future acquisition of data cheaper and less time-consuming [10–11].

The aim of this paper is to discuss different variable selection procedures for multi-block regression data. In particular, the selection of variables will here be coupled with MB-PLS [4,12] and SO-PLS [6,13] models, which are both based on PLS regression. A simulation study will be conducted for regular one-block PLS regression, in order to select which variable selection methods to include in the multi-block study. Details on this simulation are reported in Appendices A and B. Three candidate variable selection methods will be used in order to obtain insight into the influence of the choice of the variable selection method. The different procedures will be illustrated with different data sets; one sensory data set with relatively few samples and variables, one spectroscopic data set with more samples and many correlated variables and a number of simulated multi-block data sets.

## 2. Multi-block methods

In this section, we present the multi-block methods applied in the paper and also an overview of the procedures used for implementing variable selection. A more detailed discussion of the choice of the actual PLS variable selection methods to be used within MB-PLS and SO-PLS is given in Appendices A and B. Only one  $Y$ -variable and two input blocks are considered here, but the multi-block methodology can easily be extended. In this paper we will assume the linear model structure:

$$Y = Xf + Zg + E \quad (1)$$

where:  $X$  ( $N \times J$ ) and  $Z$  ( $N \times L$ ) are the predictor blocks and  $Y$  ( $N \times 1$ ) is the response variable.  $E$  ( $N \times 1$ ) is the residual matrix and  $f$  and  $g$  are

\* Corresponding author at: Nofima AS, Osloveien 1, P.O. Box 210, N-1431, Ås, Norway.  
E-mail address: [alessandra.biancolillo@nofima.no](mailto:alessandra.biancolillo@nofima.no) (A. Biancolillo).



the regression coefficients of dimension  $(L \times 1)$  and  $(J \times 1)$ , respectively. All variables are assumed to be mean centered.

### 2.1. Multi-Block-PLS regression

The Multi-Block-PLS method (MB-PLS) [4,12] is based on concatenating the input blocks and then performing PLS regression on the resulting matrix  $\mathbf{X}_{conc}$ . In general, the matrices are block-scaled before concatenation. Block-scaling can be performed in different ways; the one pursued in this work is based on dividing each block by its Frobenius norm. This scaling aims to ensure that no block will be more dominant than others because of the number of variables and their variance.

### 2.2. Sequential and Orthogonalized Partial Least Squares regression

Sequential and Orthogonalized Partial Least Squares (SO-PLS) [6,13] is a multi-block method that in the case of two blocks can be described as follows:

1.  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  by PLS-regression
2.  $\mathbf{Z}$  is orthogonalized (obtaining  $\mathbf{Z}_{orth}$ ) with respect to the scores of the previous PLS model
3.  $\mathbf{Y}$  residuals from the first PLS are fitted to  $\mathbf{Z}_{orth}$
4. The full predictive model is computed by summing up the two contributions from  $\mathbf{X}$  and  $\mathbf{Z}$ .

If more than two predictor blocks are involved, it is possible to perform SO-PLS repeating the steps, as explained in [13]. The optimal complexity is estimated from the so-called Måge-plot as described in [13]. Two different approaches can be chosen: *global optimization* and *sequential optimization*. The strategy pursued here is the former one.

The SO-PLS method is invariant to block scaling and explicitly permits the interpretation of the contributions of the blocks and their relationship with the response. It can also be used to handle blocks with very different underlying dimensionality, such as for instance design variables and multivariate spectra, in the same model. The  $\mathbf{X}$ -block is interpreted by inspecting the PLS model in step 1. The interpretation of the  $\mathbf{Z}$ -block is best done by calculating loadings by projecting  $\mathbf{Z}$  onto the scores obtained in step 3 [14].

## 3. PLS variable selections methods

There are many methods for variable selection in general and for PLS in particular [15–18,20–26]. For the purpose of doing a sensible multi-block variable selection, we tested a number of established PLS variable selection methods in a preliminary simulation study (Appendices A and B). Based on the results, two candidate methods were selected to be used in the different PLS based multi-block models. These are *Variable Importance in Projection* (VIP) and *Selectivity Ratio*. In addition to these two, *forward selection* was also included for comparison. More details about these choices can be found in Appendix B, together with a description of all the tested methods, details on the ANOVA used and main results.

### 3.1. Variable selection for multi-block methods

In the following, we will describe different procedures for combining variable selection with MB-PLS and SO-PLS. In particular, we will focus our discussion on:

- 1) MB-PLS combined with VIP
- 2) MB-PLS combined with SR
- 3) SO-PLS with pre-selected variables using VIP on each block
- 4) SO-PLS with pre-selected variables using SR on each block
- 5) SO-PLS combined with VIP
- 6) SO-PLS combined with SR
- 7) SO-PLS combined with forward selection.

All the different procedures are described below and summarized in Table 1. We will refer to blocks  $\mathbf{X}$  and  $\mathbf{Z}$  after variable selection as  $\mathbf{X}_{Red}$  and  $\mathbf{Z}_{Red}$ .

#### 3.1.1. Proposed procedure for variable selection in MB-PLS

The selection of variables in MB-PLS is an issue that has not yet been explored, although a reinterpretation of MB-PLS as a variable selection method itself [19] has been suggested. The procedure proposed in this paper (points 1 and 2 in the list at the beginning of Section 3.1) is to perform variable selection (using SR or VIP) directly on the concatenated input matrix. Following the standard MB-PLS procedure, predictor blocks are block-scaled, concatenated, and then PLS is performed on the resulting matrix  $\mathbf{X}_{conc}$ . Variable selection is then based on the obtained PLS model. This leads to a number of variables being selected and  $\mathbf{X}_{conc}$  is reduced obtaining  $\mathbf{X}_{Red}$ . Finally, a new calibration model is obtained for  $\mathbf{Y}$  using the reduced matrix  $\mathbf{X}_{Red}$  in a new MB-PLS model.

#### 3.1.2. Proposed procedures for variable selection in SO-PLS

One possible approach in SO-PLS is to select variables from each block separately (points 3 and 4 in the list in Section 3.1). In other words,  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  and to  $\mathbf{Z}$  independently, creating two different PLS models. Variables in each block are selected (by SR or VIP) and the two sets of variables are then used in SO-PLS. Note that it is possible to leave one of the blocks untouched; i.e. to perform variable selection on only one of the blocks. When selection is done on both,  $\mathbf{X}_{Red}$  and  $\mathbf{Z}_{Red}$  are obtained and used in the SO-PLS regression. Compared to the procedure in Section 3.1.1, however, there is a risk of overlooking possible synergies between the blocks with this approach.

An alternative is to integrate variable selection directly into the SO-PLS algorithm (points 5 and 6 in the list in Section 3.1). Due to the sequential nature of the SO-PLS method, variables can be selected (by VIP or SR) from the  $\mathbf{X}$ -block, from the  $\mathbf{Z}_{orth}$ -block or from both. When the variable selection involves both blocks, the algorithm is the following:

1.  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  by a PLS model.
2. A variable selection method is applied to  $\mathbf{X}$  obtaining  $\mathbf{X}_{Red}$ .
3.  $\mathbf{Y}$  is refitted to  $\mathbf{X}_{Red}$ .
4.  $\mathbf{Z}$  is orthogonalized with respect to the scores of the PLS model in step 3.
5. The residual matrix from step 3 is fitted to  $\mathbf{Z}_{orth}$ .
6. A variable selection method is applied to  $\mathbf{Z}_{orth}$  obtaining  $\mathbf{Z}_{orth,Red}$ .
7. A new PLS regression is carried out using the reduced matrix  $\mathbf{Z}_{orth,Red}$  to fit the residual matrix.
8. The full predictive model is computed by combining the contributions in the same way as in the original model.

When the variable selection involves only one block, the steps related to the reduction of variables in the other block (steps 2 and 3 or 6 and 7) are skipped. When for instance only the  $\mathbf{X}$ -block is reduced, the model will coincide with the one built from SO-PLS using  $\mathbf{X}_{Red}$  and  $\mathbf{Z}$  in the previous procedure.

After the reduction of the blocks the model is rebuilt using the reduced blocks and the optimal number of latent variables is redefined on the reduced blocks by means of the Måge plot. The algorithm is forced to select at least one latent variable for each block. Hence, solutions that do not select any latent variables in one of the two blocks are skipped.

**Table 1**  
Combined multiblock and variable selection methods.

Variable selection method	Multiblock method	
	MB-PLS	SO-PLS
VIP	✓	✓
SR	✓	✓
Forward selection		✓

The final proposed procedure to perform variable selection in SO-PLS is an extension of the forward selection method (point 7 in the list at the beginning of Section 3.1).

First, the best predictor is selected from either  $\mathbf{X}$  or  $\mathbf{Z}$  based on RMSECV. Next, each of the successively added variables will come either from  $\mathbf{X}$  or  $\mathbf{Z}$ . The algorithm will test all the possible combinations which result from either adding one variable from the  $\mathbf{X}$ -block keeping the  $\mathbf{Z}$ -block as in the previous step or vice versa. At the  $v + 1$ th iteration,  $v_1$  and  $v_2$  variables (with  $v_1 + v_2 = v$ ) have already been selected from the  $\mathbf{X}$ - and the  $\mathbf{Z}$ -blocks, respectively. Then, the algorithm proceeds by building  $J - v_1$  SO-PLS models, considering all the possible combinations resulting from the addition of one more  $\mathbf{X}$ -variable. Likewise,  $L - v_2$  SO-PLS models are built adding one further  $\mathbf{Z}$ -variable to  $\mathbf{Z}_{red}$  (retaining only the previously selected  $v_1$  predictors in the  $\mathbf{X}$ -block). The combination that results in the lowest RMSECV is selected. The procedure is then repeated for the selection of further variables. It is stopped when the addition of another predictor does not significantly improve the RMSECV (the significance of the addition is checked by CVANOVA [27] with a confidence level of 95%). Here it must be stressed that, if in the initial iterations all the selected variables come from a single block, the effect of the addition of a further variable to that block is tested effectively using PLS instead of SO-PLS.

Note that this method is very time consuming when the number of variables is large. However, it can be speeded up to handle for instance spectroscopic intervals instead of individual variables [28]. This will be applied to the spectroscopic data set below. Using intervals on, e.g. spectroscopic data not only speeds up the algorithm, but can also minimize overfitting tendencies which is a danger for all variable selection methods.

#### 4. Data sets

The different proposed procedures (described above in Section 3.1) have been tested on simulated multi-block data sets and on two real data sets, a spectroscopic one (Raman) and a sensory data set.

##### 4.1. Simulated multi-block data sets

Six different multi-block data sets were simulated. In all data sets, the  $\mathbf{X}$ - and  $\mathbf{Z}$ -blocks have the same number of objects (two hundred) but different numbers of variables. Variables are divided into 'selective', 'relevant but not selective', systematic but 'irrelevant' and noise variables. Those called 'selective' are only related to the response, while the 'irrelevant' variables are not. The 'relevant but not selective' ones contain information about both 'selective' and 'irrelevant' variability. Finally, some noise variables are randomly generated. The structure of this data set resembles the one used for the simulations used for selecting the most appropriate PLS variable selection method and therefore important details can be found in Appendix A. The different dimensions of the blocks are reported in Table 2.

For all the data sets, the  $\mathbf{X}$ -block is simulated by multiplying randomly generated scores ( $\mathbf{T}_X$ ) and loadings ( $\mathbf{P}_X$ ). Both scores ( $\mathbf{T}_X$ ) and loadings ( $\mathbf{P}_X$ ) are simulated from the normal distribution  $N(0,1)$ .

The  $\mathbf{T}_X$  has fixed dimensionality ( $200 \times 4$ ) where only the first three components are 'selective'.  $\mathbf{P}_X$  is a partitioned matrix constructed to reflect the fact that there are variables in the four different categories 'unique-selective', 'unique-irrelevant', 'relevant but not selective' and noise. (For details regarding score and loading structures, please look at the simulated Dataset-1 described in Appendix A. The  $\mathbf{X}$ -block here is generated following the same procedure used for the generation of  $\mathbf{X}$  in the simulation presented in Appendix A).

The  $\mathbf{Z}$  scores are correlated with  $\mathbf{X}$ .  $\mathbf{Z}$ -scores  $\mathbf{T}_Z$  are divided into  $\mathbf{T}_{Zsel}$  ( $200 \times 2$ ) and in  $\mathbf{T}_{Zirr}$  ( $200 \times 1$ ).  $\mathbf{T}_{Zsel}$  is a partitioned matrix of the form:

$$\mathbf{T}_{Zsel} = [\mathbf{T}_{Z1} \ \mathbf{T}_{Z2}] \quad (2)$$

where  $\mathbf{T}_{Z1}$  ( $200 \times 1$ ) is a linear combination of the first two columns of  $\mathbf{T}_X$  and  $\mathbf{T}_{Z2}$  ( $200 \times 1$ ) containing random values drawn from the normal distribution  $N(0,1)$ .

The  $\mathbf{Z}$  loading matrix ( $\mathbf{P}_Z$ ) is a partitioned matrix (as  $\mathbf{P}_X$ ) reflecting the four different categories of variables. The data matrices  $\mathbf{X}$  and  $\mathbf{Z}$  are generated as  $\mathbf{X} = \mathbf{T}_X \mathbf{P}_X^T$  and  $\mathbf{Z} = \mathbf{T}_Z \mathbf{P}_Z^T$  and subsequently,  $\mathbf{Y}$  is calculated as:

$$\mathbf{Y} = [\mathbf{T}_{Xsel} \ \mathbf{T}_{Zsel}] * \boldsymbol{\beta} \quad (3)$$

The vector  $\boldsymbol{\beta}$  ( $5 \times 1$ ) is generated as a matrix containing random values drawn from the uniform distribution (mean is 0.55) in the open interval (0.05, 1.05). The  $\mathbf{Z}$ -Loadings and test sets are generated as in Dataset-1 (see Appendix A).

Finally, random noise corresponding to 10% of the signal was added to all the predictors of the data sets. For the responses, the added noise corresponded to 5% of the signal.

As shown in Table 2, four data sets (*Sim1*, *Sim2*, *Sim3* and *Sim4*) have comparable amount of samples and variables. Instead, the last two (*Sim5* and *Sim6*) have blocks with more variables than objects.

Each data set was generated one hundred times. All the proposed variable selection procedures for multi-block data have been tested on all the training sets. Test sets were generated in the same way as the training data but with 300 samples. Reduced test sets were then obtained (taking only the variables that were selected on the training sets) and used for the validation. It is important to stress that the test sets were not involved in the selection of the variables. Test sets are reduced after the selection is done on the training sets, then they are used to perform the external validation.

##### 4.2. Flavored waters data set

The data set is based on sensory analysis and consumer liking of eighteen flavored waters [6]. The purpose is to get insight into which sensory attributes that are most related to consumer liking. Samples have been recorded based on a full factorial design. Three factors are taken into account: flavor type (A and B), sugar dose (2%, 6% and 8%) and flavor dose (Low, Medium and High). This gives 18 samples in total. Eleven trained assessors evaluated samples by smelling and

**Table 2**  
Parameters used for the generation of the six different simulated multiblock datasets.

Simulation	$\mathbf{X}$ -block			$\mathbf{Z}$ -block		
	# Selective variables	# Relevant but non-selective variables	# Irrelevant variables	# Selective variables	# Relevant but non-selective variables	# Irrelevant variables
<i>Sim1</i>	30	30	30	40	40	40
<i>Sim2</i>	50	20	20	60	20	40
<i>Sim3</i>	80	20	20	60	30	0
<i>Sim4</i>	120	0	0	100	0	30
<i>Sim5</i>	350	100	50	300	100	50
<i>Sim6</i>	350	100	100	300	100	0

tasting. The evaluation of the smell attributes resulted in the **Smell**-block, while the evaluation of the *taste attributes* constitute the **Taste**-Block (see Table 3).

The smell data are used in the following analysis as the **X**-block, and the taste as **Z**-block. A major interest in this setup is to assess how much extra information about liking one obtains by adding taste to the smell variables. All sensory data used here were averaged over assessors. Finally, the consumers' rating of the waters (ranked from 1-“Dislikes very much” to 9-“Likes very much”) are collected. The average rates over the consumers are used as response.

### 4.3. Polyunsaturated fatty acids (PUFA) data set

Sixty-nine emulsions of defatted whey protein concentrate, water, and five different oils, (*olive oil, coconut oil, soy oil, cod oil enriched with polyunsaturated omega-3 fatty acids, and salmon oil*) were analyzed by Raman spectroscopy. Each sample represents a different amount of the various constituents. These amounts were defined based on an experimental design; more details can be found in [29]. The Raman spectra have been divided into two blocks. One block is the one containing the so-called *Fingerprint region* (wavelengths from 675 to 1197 cm<sup>-1</sup>), and is the one used as the **X**-block in the analysis. The relevance of the fingerprint region is that each compound produces a characteristic pattern in this part of the spectrum. Therefore, it is relevant to investigate this data block separately and together with the remaining spectral information. The second block is constituted by spectra from 1198 to 1770 cm<sup>-1</sup> and is used as the **Z**-block. This is the region of the spectrum where the main absorptions of the functional groups of each compound take place. Concentrations of PUFA in the emulsions are used as response.

### 4.4. Data analysis

All data analyses were performed using MATLAB (R2012b, The Mathworks, Natick, MA), using in-house routines for PLS, MB-PLS, SO-PLS and for all the variable selection methods. The MATLAB routines for MB-PLS and SO-PLS are available for download at [www.nofimamodeling.org](http://www.nofimamodeling.org).

## 5. Results

All the proposed procedures discussed in Section 3.1 have been applied to the multi-block simulated data sets and to the real data sets. Selectivity Ratio has been applied using two different cut-off values: one based on the *F*-test and one based on its mean (see Appendix B for details). Results obtained using both cut-off values are reported for the sensory data set. For the simulated multi-block data sets only the cut-off based on the mean was used. Instead, only

**Table 3**  
Sensory descriptors in the flavored waters data set. Numeration of variables is reported to help the comprehension of the discussion in Section 5.

Var. number	Smell	Var. number	Taste
1	Ripe	1	Ripe
2	Tropical	2	Tropical
3	Candy	3	Candy
4	Synthetic	4	Synthetic
5	Lactonic	5	Lactonic
6	Sulfuric	6	Sulfuric
7	Skin	7	Skin
8	Green	8	Green
9	Floral	9	Floral
		10	Sweet
		11	Sour
		12	Bitter
		13	Dry
		14	Sticky

the *F*-test based cut-off was applied for the spectroscopic data set. The reasons for these choices are reported in the relevant subparagraphs.

### 5.1. Simulated multi-block data sets

The predictive ability of the models was assessed by the external test set using the Root Mean Square Error of Predictions (RMSEPs). The selected variables for two different data sets (*Sim1* and *Sim5*) are reported in Table 4. Results from the other data sets are in agreement with these, both for predictions and interpretations and are therefore not shown in detail.

Variable selection tends to improve the predictions compared to the full models. The best predictions, both in *Sim1* and in *Sim5*, are obtained using SR in combination with SO-PLS.

From an interpretation point of view, the results are similar to what was observed in the simulation study of regular PLS regression reported in Appendix B. SR retains its ability in skipping almost all the ‘irrelevant’ variables. In fact, it does not select any ‘irrelevant’ variable when applied to MB-PLS. The SR behaves differently when implemented in procedure 6 and when applied in procedure 4 (from the list in Section 3.1). In procedure 6, it selects few ‘irrelevant’ variables from the **X**-block and none from the **Z**-block (in both *Sim1* and *Sim5*). When applied in procedure 4, it also selects few ‘irrelevant’ variables (2% and 4% in *Sim1* and *Sim5*, respectively) from the **X**-block, but many from the **Z**-block (34% and 35% for both data sets). SR combined with SO-PLS selects several ‘selective’ variables, but always fewer than when using VIP. Moreover, SR selects fewer or a comparable number of ‘relevant but not selective’ variables than VIP when selecting from the **X**-block, but more than VIP when it comes to the **Z**. Furthermore, SR is good at removing noise variables. In conclusion, SR is best in skipping unrelated information when it is integrated into the model (procedure 6) and not done beforehand on the individual blocks (procedure 4).

VIP is good at selecting the ‘selective’ variables. Looking at Table 4, it is always selecting many ‘selective’ variables from both blocks and in both data sets. It also selects a high number of ‘relevant but not selective’ ones. In accordance with to the simulation in Appendix B.3, it skips completely the noise variables. Consequently, VIP is suggested for selecting the relevant information in multi-block data sets, independent of the regression method used to handle them. Additionally, it is recommended for noisy multi-block data sets.

Applying the forward selection to SO-PLS, a rather small number of ‘selective’ and ‘relevant not selective’ variables is selected. Here this method does not show any particular ability in skipping the ‘irrelevant’ variables and the noise.

In conclusion, from the simulated multi-block data sets, it appears that SR, in general, is able to eliminate noise variables. It selects a substantial number of ‘selective’ and ‘relevant but not selective’ variables from **X** and it selects more ‘relevant but not selective’ variables than VIP from **Z**. It skips completely the ‘irrelevant’ variables when combined with MB-PLS and when implemented in SO-PLS. Moreover, when SR is combined with SO-PLS (both procedures 4 and 6 in Section 3.1) the lowest RMSEPs are obtained. VIP selects many ‘selective’ and ‘relevant but not selective’ variables, and it is efficient in skipping the noise variables. It is the recommended method for noisy multi-block data sets when the main aim is interpretation. Forward selection selects some ‘relevant’ variables, but also ‘irrelevant’ and noisy ones.

### 5.2. Flavored waters data set

Since the flavored waters data set has a limited number of samples it was not possible to have an external validation set. Therefore, all the models are cross-validated (by leave-one-out cross-validation). The prediction results for all the different methods described in Section 3 are reported in Table 5. SR was applied using both the *F*-test and SR's mean as cut-off values (see Appendix B for details). From the prediction point of view, results obtained using the two different cut-off values are

**Table 4**

RMSEPs for the prediction of  $y$  (from the simulated multiblock datasets *Sim1* and *Sim5*) by PLS, MS-PLS, SO-PLS and by MS-PLS and SO-PLS combined with variable selection methods. Relative percentages of the different type of variables selected from the procedures are also reported.

Procedure	Variable selection method	Var. selected in <i>X</i> -block (%)				Var. selected in <i>Z</i> -block (%)				RMSEP
<i>Sim1</i>										
		Sel (%)	ReInSel (%)	Irr (%)	Noise (%)	Sel (%)	ReInSel (%)	Irr (%)	Noise (%)	
MB-PLS	No var. sel	All	All	All	All	All	All	All	All	1.22
	VIP	85	75	16	0	68	48	31	0	1.12
	SR	20	24	0	0	41	71	0	0	1.11
Selection on individual block + SO-PLS	No var. sel	All	All	All	All	All	All	All	All	
	VIP	78	64	28	0	74	50	34	0	0.79
	SR	45	59	2	0	14	70	34	0	0.64
Selection integrated in SO-PLS	No var. sel	All	All	All		All	All	All		
	VIP	80	64	28	0	72	47	45	0	0.80
	SR	45	59	2	0	11	73	0	0	0.59
	Forw. sel.	16	3	4	23	19	1	9	8	0.64
<i>Sim5</i>										
MB-PLS	No var. sel	All	All	All		All	All	All		1.21
	VIP	49	10	4	0	59	14	27	0	1.25
	SR	17	6	0	0	37	23	0	0	1.14
Selection on individual block + SO-PLS	No var. sel	All	All	All		All	All	All		
	VIP	47	12	24	0	56	13	36	0	0.66
	SR	34	13	4	0	13	24	35	0	0.59
Selection integrated in SO-PLS	No var. sel	All	All	All		All	All	All		
	VIP	47	12	24	0	56	14	41	0	0.69
	SR	34	13	4	0	11	25	0	0	0.59
	Forw. sel.	8	6	12	40	6	2	14	10	0.70

comparable. Concerning the interpretation, the main difference is that a different number of variables (in particular in the second block) are selected. In the discussion below, when not stated differently, we are referring to SR with  $F$ -test as cut-off value.

As can be seen from Table 5, the RMSECV obtained from PLS on the smell block alone is comparable to those obtained by the multi-block approaches, meaning that from a prediction point of view the taste block adds little information. The only substantial improvement in RMSECV is given by SO-PLS using forward selection as variable selection method. A possible reason for this could be that it selects variables according to predictive ability and is then more sensitive to overfitting, especially for such a small data set. But it could also be an indication of real improvement. However, it is still of interest to apply variable selection using a multi-block approach, for the sake of interpretation.

In most models, SR selects more variables than VIP in  $X$ , but when it comes to  $Z$  it depends on the procedure used. Variable selection by SR does not select  $Z$ -variables when applied in MB-PLS. Concerning SO-PLS, the number of selected variables in each blocks depends on when the variables are selected. If variables are selected on the individual blocks before creating the SO-PLS model (procedures 3–4 from the list in Section 3.1), VIP selects more  $Z$ -variables than SR; when it is

implemented in the SO-PLS building (procedures 5–6 from the list in Section 3.1), it is the other way around. When SR is applied for the individual blocks before building the SO-PLS model (procedure 4), it selects just one variable. In the preliminary PLS study (Appendix B.3), SR shows a good ability to not select ‘irrelevant’ variables. That suggests that  $Z$ -variables could be considered ‘irrelevant’, confirming the results above that the taste block is not adding much to the predictive ability of models. The situation is quite different when SR is integrated into the SO-PLS model. This is probably due to the fact that, in this case, variables are not selected directly on the  $Z$ -block, but on  $Z_{orth}$ . One of the drawbacks of the orthogonalization in SO-PLS is that, after the first regression, some of the noise goes into the residuals. Residuals are then fitted to  $Z_{orth}$ ; consequently, noisy data can affect this part of the modeling.

In simulations, VIP has demonstrated a better ability to handle the noise than SR. This explains why the number of variables selected from  $Z$  by VIP when combined with SO-PLS is quite the same (three when the selection is done beforehand and four when it is implemented in the SO-PLS), while SR behaves differently (one variable when the selection is done beforehand and nine when it is implemented in SO-PLS).

**Table 5**

RMSECVs and explained variance for the prediction of  $y$  (Sensory dataset) by PLS, MS-PLS, SO-PLS and by the different variable selection procedures in multiblock ( $X$ -block: Smell-block;  $Z$ -block: Taste-block). Selected variables from the different methods and number of variables used in each model are also reported.

Procedure	Variable selection method	Selected variables smell	Selected variables taste	LVs	RMSECV	Explained variance $Y$ (%)
No variable selection	Only smell (PLS)	All	None	1	0.25	53
	Only taste (PLS)	None	All	1	0.33	18
	MB-PLS	All	All	1	0.26	50
	SO-PLS	All	All	1,1	0.26	48
MB-PLS	VIP	1; 2; 4; 5; 6	4	1	0.26	50
	SR	1; 2; 4; 5; 6; 8	None	1	0.25	54
	SR <sub>(mean)</sub>	1; 4; 5; 6; 8	None	1	0.25	54
	VIP	1; 4; 5; 6	1; 4; 5; 9	2, 1	0.24	56
Selection on individual block + SO-PLS	SR	1; 2; 4; 5; 6; 8	4	1, 1	0.23	60
	SR <sub>(mean)</sub>	1; 4; 8	1; 2; 4; 5; 6	1	0.26	48
	VIP	1; 4; 5; 6	1; 4; 10	2, 1	0.24	55
Selection integrated in SO-PLS	SR	1; 2; 4; 5; 6; 8	1; 2; 4; 7; 8; 10; 11; 13; 14	1, 1	0.25	53
	SR <sub>(mean)</sub>	1; 4; 8	1; 7; 10; 11; 13	1	0.28	48
	Forw. sel.	2; 3; 6	8	1, 1	0.21	66



For VIP, it is quite consistent in its selection on  $\mathbf{X}$ , independently of the method/model. VIP always selects variables number 1, 4, 5, 6 (ripe, synthetic, lactonic and sulfuric, respectively). When applied to MB-PLS it also selects variable number 2, tropical. On the  $\mathbf{Z}$ -block the selection is less consistent, but variable number 4 (synthetic) is always selected.

This data set is useful for investigating how SO-PLS handles a multi-block set since it has the interesting characteristic of having the first nine attributes in common in the two blocks. Fig. 1 shows the selected variables in the two blocks when Both VIP and SR are integrated into the SO-PLS model. Fig. 1(a) shows the selected variables by VIP and Fig. 1(b) those selected by SR. For VIP, it seems that the relevant variables belong mainly to the “common” ones (same attributes for smell and taste). In fact, when VIP is used to select variables, only one “unique feature” (an attribute not present in both blocks) is selected in the Taste-Block (number 10, Sweet). SR is less parsimonious and selects four of the variables that belong only to the Taste-Block.

In SO-PLS we expect that the common variation between  $\mathbf{X}$  and  $\mathbf{Z}$  is explained by  $\mathbf{X}$  and then removed from the  $\mathbf{Z}$ -Block. Therefore, smell variables that are selected in the  $\mathbf{X}$ -Block are not expected to be selected again as taste variables in the  $\mathbf{Z}$ -block. As can be seen in Fig. 1, some common variables are selected from both blocks in this example. The reason for this is likely that the same attributes are sometimes perceived differently when tasting, so even if they have the same name, the correlation between smell and taste might be low. This is for instance the case for variable 1 (ripe), 2 (tropical), 4 (synthetic), and 8 (green), which are selected from both smell and taste with the SR method. The correlation between smell and taste for these attributes are 0.6, 0.6, 0.7 and 0.4 respectively. On the contrary, variable 5 (lactonic) and 6 (sulfuric) are selected only from the smell block. They both have correlation 0.8, indicating that the attributes are perceived similarly by tasting and smelling. Variable number 7 (skin), on the other hand, is only selected from the taste block. For this attribute, the correlation is actually zero, and hence it is a completely different perception in the taste block. In addition, we noticed that all the variables selected by both blocks have a higher SR value in  $\mathbf{X}$  than in  $\mathbf{Z}$ . This means that the variation that is “left” in  $\mathbf{Z}$  is less important, since some of it is already accounted for by  $\mathbf{X}$ .

The forward selection approach is extremely focused on selecting only *non-common* variables between the predictors. As shown in Fig. 2, there is no overlap between the selected variables in the two blocks.

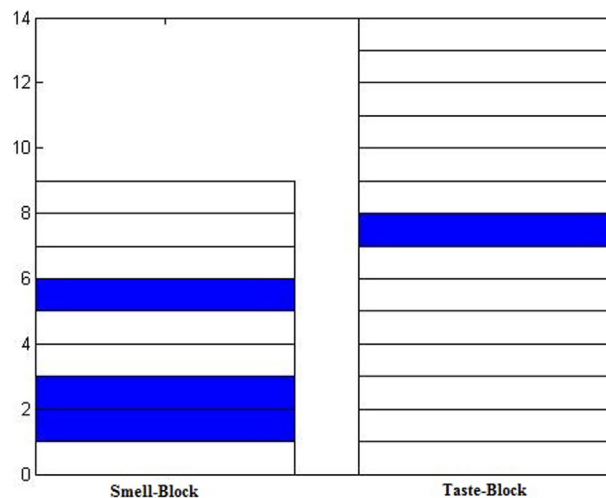


Fig. 2. Selected variables by the forward selection combined with SO-PLS. Selected variables are highlighted in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 5.3. Results on the PUFA data set

The PUFA data set was split into training and test sets (by the Duplex algorithm [30]) in order to use the latter for validation. Forty-nine samples were selected for the training set, while the test set is composed of twenty samples. The training set was used to select variables, build different calibration models and select number of components. The test set was then used for calculating RMSEP. Results are reported in Table 6. For SR, the cut-off value used is the one based on the  $F$ -test. Also the other cut-off value was tested, but led to worse predictions. Therefore, it is not mentioned further in the following. From Table 6 one can see that 96% of the variation in the response is explained by  $\mathbf{Z}$  alone, and combining  $\mathbf{X}$  and  $\mathbf{Z}$  does not improve the prediction ability much. This means that also in this case, the main motivation for doing multi-block analysis is interpretation.

In order to perform the forward selection on the spectroscopic data set, the training set (both  $\mathbf{X}$  and  $\mathbf{Z}$ ) is divided into 20 intervals (with approximately the same number of variables for each interval belonging to the same block), and then the forward selection is applied as

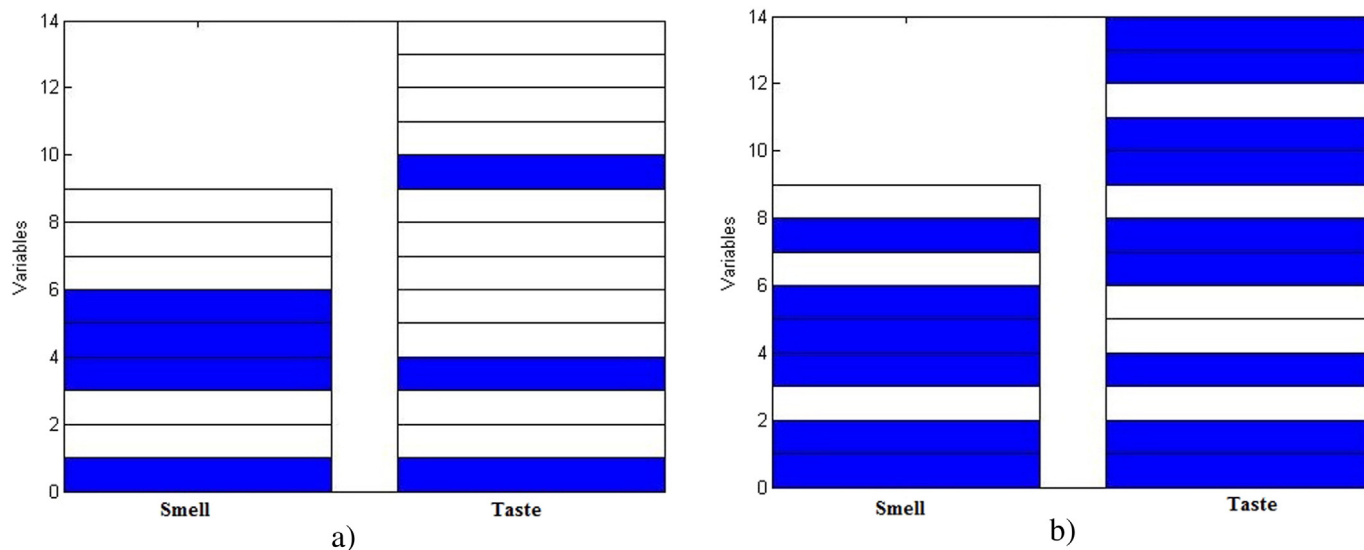


Fig. 1. Selected variables in  $\mathbf{X}$ - and  $\mathbf{Z}$ -blocks by variable selection integrated into SO-PLS models. Selected variables are highlighted in blue; (a) variables selected by VIP (b) variables selected by SR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 6**

RMSECVs and explained variance for the prediction of  $y$  (Raman dataset) by PLS, MB-PLS and SO-PLS in combination or not with variable selection methods. The number of selected variables from different methods and the total number of variables used in each model are also reported.

Procedure	Variable selection method	Selected variables <b>X</b>	Selected variables <b>Z</b>	LVs	REMSEP	Explained variance <b>Y</b> (%)
No variable selection	Only <b>X</b> (PLS)	All	None	3	1.61	86
	Only <b>Z</b> (PLS)	None	All	4	0.88	96
	MB-PLS	All	All	4	1.00	95
	SO-PLS	All	All	3, 3	0.90	96
MB-PLS	VIP	230/523	202/574	4	1.02	94
	SR	83/523	112/574	4	2.02	75
	SR <sub>(mean)</sub>	157/523	202/574	3	2.72	62
	Selection on individual block + SO-PLS	VIP	182/523	136/574	3, 4	1.07
SR		52/523	65/574	1, 7	0.79	97
SR <sub>(mean)</sub>		152/523	193/574	3, 2	1.16	93
Selection integrated in SO-PLS		VIP	182/523	129/574	4, 5	1.24
	SR	52/523	53/574	4, 1	1.30	95
	SR <sub>(mean)</sub>	152/523	102/574	3, 2	1.09	94
	Forw. Sel.	52/523	29/574	4, 1	1.19	94

described in Section 3.1.2, but using intervals of contiguous variables instead of individual variables. Consequently, the best combinations of intervals are selected. Three intervals in total gave the lowest RMSECV; two interval for the **X**-block and one interval from the **Z**-block. This amounts to 52 variables from the **X**-block and 29 from the **Z**-block.

As can be seen from Table 6, the number of variables is strongly reduced by all methods but, as opposed to the flavored waters example, the VIP method consistently selects 2–3 times more variables than SR in both **X** and **Z**, regardless of the variable selection method.

Looking more into the selected variables, VIP and SR select different variables from the two blocks. In Fig. 3 one can see which variables were selected by VIP, SR and forward selection when integrated into the SO-PLS model. Fig. 3(a) and (b) represents spectra in **X** and **Z**, respectively. The upper curves are the average spectra (offset to make them more visible) where selected variables by SR are presented in boldface. In the middle line, the boldface variables are those selected by VIP. The lines at the bottom (offset downwards) show in boldface the variables selected by the forward selection. From the interpretation point of view, VIP is the more interesting. Indeed, looking at the fingerprint region, (Fig. 3a, middle line), it selects areas related to the skeletal C–C, C–N and to the C–O stretching (1080, 1060, 925, and 864  $\text{cm}^{-1}$ ). For the **Z**-block (Fig. 3b), VIP is able to select the most relevant bands. In fact, selected variables are those around 1263  $\text{cm}^{-1}$ , where the symmetric rocking of  $=\text{C}-\text{H}$  takes place. Moreover, it selects variables around 1445  $\text{cm}^{-1}$  where the  $\text{CH}_2$ 's scissoring takes place, and variables

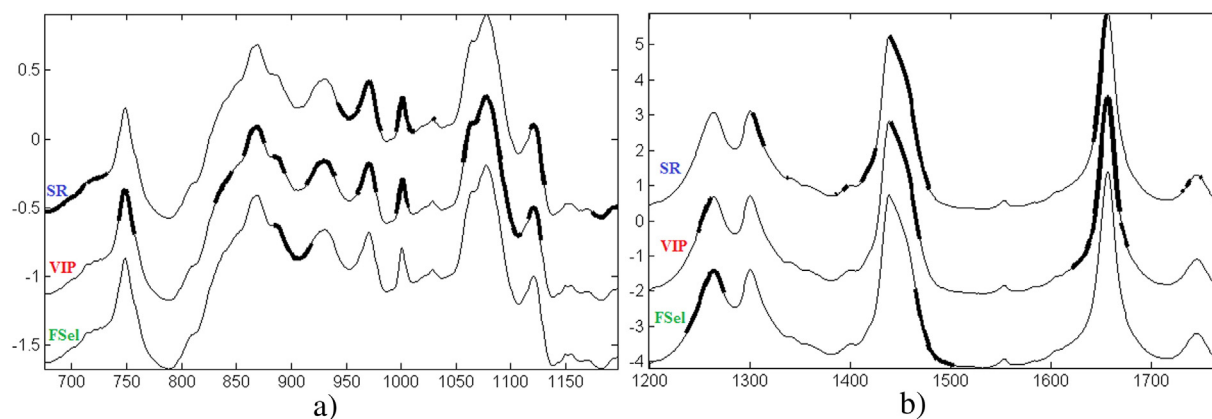
around 1656  $\text{cm}^{-1}$  where there are the  $\text{C}=\text{C}_{(\text{cis})}$  stretching and amide I absorptions.

When variable selection is done beforehand on the individual blocks (procedures 3 and 4 in the list in Section 3.1), VIP selects only seven variables more than those selected following the procedure 5 in Section 3.1.

The differences in the behavior of SR and VIP can be explained from the results of the simulation studies. Here, it is evident that some wavelengths selected from VIP are not selected by SR (in particular on the **X**-block). Since the Raman spectra are measurements of mixtures of water, whey proteins and oils, this finding could be due to the fact that not only PUFA is contributing to the Raman signal. Some wavelengths are related to functional groups present both in PUFA and in whey proteins. These variables are 'relevant but not selective' (because they are not univocally related to the PUFA). As observed in the simulation study in Appendix B.3, SR selects less 'relevant but not selective' variables than VIP. Consequently, the behaviors observed are not surprising.

Predictions made without variable selection are similar to those obtained by reduced models. This could be taken as an indication that the presence of the whey proteins has, at best, a moderate additional effect on the spectroscopic signal.

The forward selection applied to SO-PLS gives less interesting results than VIP from the interpretation point of view. It selects many seemingly relevant peaks but some are also missed out. Concerning the fingerprint part of the Raman spectrum (Fig. 3(a), bottom line), it selects



**Fig. 3.** Selected variables when VIP, SR and forward selection are implemented in SOPLS (procedures 5–7). (a) Lines represent the average spectra in **X**. The upmost lines are the average spectra (offset to make them more visible) where the selected variables by SR are bolded. The middle lines are average spectra (offset downwards) where the bolded variables are those selected by VIP. The lowest lines are average spectra (offset downwards) where the bolded variables are those selected by the forward selection (b) corresponding plot for **Z**.

variables related to the skeletal C–O stretching (around  $925\text{ cm}^{-1}$ ). Looking at the rest of the Raman spectra (Fig. 3(b), bottom line) it picks the  $\text{CH}_2'$  twisting and the  $=\text{CH}'$  symmetric rocking (variables between  $1200$  and  $1356\text{ cm}^{-1}$ ).

## 6. Discussion and conclusions

In the present paper, different approaches for performing variable selection in a multi-block context have been proposed. All the proposed procedures conceived for selecting variables in the framework of MB-PLS and SO-PLS were tested on different simulated data sets and on two real ones.

Below we present some suggestions for selecting an appropriate approach for variable selection in multi-block regression. The results are also summarized in a flow chart in Fig. 4.

### 6.1. Prediction

Inspecting the simulated multi-block data sets, it appears that SO-PLS combined with any of the proposed variable selection methods (also the SO-PLS in itself) gives models with good predictions. In particular, SO-PLS (with or without variable selection) performs better than the MB-PLS models. Predictions are particularly good when SO-PLS is combined with SR.

It has to be highlighted that, from a practical point of view, the effort required by selection methods based on the evaluation of parameters (*filter methods* [11]) is different from the effort required by methods that need the rebuilding of the model every time one variable is removed/added. Consequently, among all the variable selection method used in this study, the forward selection method is definitely the most computational demanding. Moreover, it has to be taken into account that, since it selects variables in accordance with the predictive capability, the forward selection can be more sensitive to overfitting when a double validation is not adopted.

### 6.2. Interpretation

In general, the interpretation of MB-PLS models (when no variable selection method is involved) is not straightforward. For SO-PLS, the interpretation of the blocks can be done investigating the  $\mathbf{X}$ - and  $\mathbf{Z}_{\text{orth}}$ -PLS-scores and loadings [13,14]. After  $\mathbf{Y}$  is fitted to  $\mathbf{X}_{\text{Red}}$ ,  $\mathbf{Z}$  is orthogonalized with respect to the scores of this regression. Consequently,  $\mathbf{Z}_{\text{orth}}$  only contains information not present in  $\mathbf{X}_{\text{Red}}$ . Interpreting the  $\mathbf{Z}_{\text{orth}}$ -PLS-scores means interpreting the  $\mathbf{Z}$ -block

without the redundant information already present in  $\mathbf{X}_{\text{Red}}$ . Since the  $\mathbf{Z}_{\text{orth}}$ -block is less complex than the  $\mathbf{Z}$ -block, it is easier to interpret.

### 6.3. Simulation study

According to the simulation studies (Appendices A and B), VIP and SR always select a large number of 'selective' variables and skip the 'irrelevant'. The main difference between VIP and SR is that SR is particularly efficient in not selecting systematic 'irrelevant' variables, while VIP does not select noise. This gives an indication of which method has to be used for handling different type of data. If the aim of the variable selection is to get rid of systematic errors, SR should be the first choice. On the other hand, handling data with many noisy variables, VIP should be preferred.

### 6.4. Sensory data set

In the sensory data set, reduced MB-PLS models and reduced SO-PLS models gave similar results, in particular regarding the selection on the Smell-block. SR is in general the most parsimonious method for selecting from the Taste-block, (except when implemented in the SO-PLS model, where various relevant variables are pointed out). Also VIP selects a modest amount of variables, both with MB-PLS and with SO-PLS.

The forward selection offers the most reduced set of selected variables but, at the same time, it gives the most different scenario. It selects three variable in  $\mathbf{X}$ ; one of these has never been selected from the other methods. Concerning the  $\mathbf{Z}$ -block, forward selection selects only one variable; this variable has been selected just once from the other procedures.

In conclusion, if the purpose of the variable selection is to obtain the most reduced set of variables possible without sacrificing the predictive ability, the forward selection combined with SO-PLS is the suggested approach. If the aim is to point out the most relevant variables, SO-PLS combined with VIP or SR is preferable.

### 6.5. Spectroscopic data

For the more collinear spectroscopy data set, it appears that the selection method used affects the results a lot. The performance of MB-PLS (when variable selection is performed by VIP) is comparable with that of SO-PLS, but with slightly poorer results from an interpretation point of view. VIP, especially when combined with SO-PLS, gives promising results in terms of chemical interpretation. When the

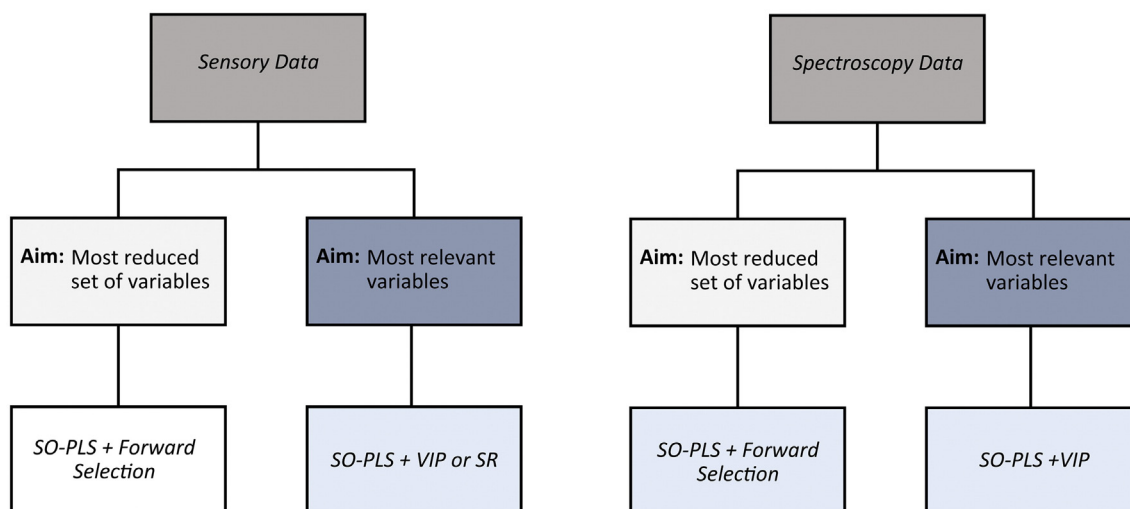


Fig. 4. Suggested variable selection approaches for sensory data and spectroscopy data.

selection is performed by this method, the most chemically-meaningful peaks are selected. SR performs parsimoniously in combination with both SO-PLS and MB-PLS to the extent that fewer chemically relevant peaks are selected.

This is a major difference between VIP and SR when applied to the sensory and to the spectroscopic data sets. When they are applied to the sensory data set, they both give good results from the interpretation point of view. When applied to the spectroscopic data set, SR misses some variables relevant for the interpretation. This may be caused by the fact that in the spectroscopic data there are more ‘relevant non-selective’ variables which SR has problems with (Section 5.3). Hence, VIP is preferred if the important variables are of this type. The forward selection gives once again the most different conclusions. It is the method that gives the most reduced set of selected variables and it skips different meaningful peaks.

In conclusion, the SO-PLS method coupled with forward selection appears to be the most preferable procedure if the focus is mainly to obtain the most reduced set of variables. On the other hand, SO-PLS in combination with VIP appear the most efficient in providing the chemical interpretation of the system. At the same time, it (VIP) also provides a reduction of the number of variables. Therefore, this is definitely the preferable approach when the focus is the exploration of the chemical meaning of the spectroscopic system.

## Acknowledgments

We would like to thank FFL (Research Levy on Agricultural Products) and the Research Council of Norway for financial support (Project number: 225096). We would also thanks to Nils Kristian Afseth, for providing the data and allowing us to use them.

## Appendices

In these appendices we present the structure and the results from the simulation conducted in order to select the most relevant PLS variable selection methods to be used together with SO-PLS and MB-PLS in a multi-block context. The multi-block simulation reported above shares several of the aspects with the structure in Appendix A and it is important for understanding the details of that simulation as well.

### Appendix A. General structure for the simulated data sets

In the first part of this work, two different data sets have been simulated in order to evaluate the power of the different variable selection methods (for PLS regression) in situations similar to the real data sets considered. The scope is to reduce the number of variable selection methods to bring into a multi-blocks PLS framework. The data sets represent an ordinary two-block regression problem, but contain several of the features of interest in a multi-block context. These same features or aspects are later on considered also in the multi-block simulation (see Section 4.1). The details on settings of the parameters are presented in Appendix B.

*Dataset-1* is created in order to mimic spectroscopic data. Therefore, the number of variables considerably exceeds the number of the samples ( $N$ ). *Dataset-2* is built with the purpose of being sensory-like in the sense that the number of columns is slightly higher than the number of rows.

Particular attention has been given to the variables’ structure from a prediction point of view. The procedure used to build *Dataset-1* is described below in detail.

*Dataset-1* is constituted of a training set ( $\mathbf{X}$  and  $\mathbf{Y}$ ) and a test set ( $\mathbf{X}_t$  and  $\mathbf{Y}_t$ ). The number of samples ( $N$ ) in the training set is defined according to an experimental design. The  $\mathbf{X}$  and  $\mathbf{Y}$  matrices have dimensions  $N \times 400$  and  $N \times 1$ , respectively. The  $\mathbf{X}$  matrix is generated as  $\mathbf{T}_x \mathbf{P}_x^T$ . The  $\mathbf{X}$ -scores  $\mathbf{T}_x$  are simulated from the normal distribution  $N(0,1)$ . The construction of  $\mathbf{P}_x$  is explained in detail below. The  $\mathbf{X}$ -block is designed

as a five-components ( $K$ ) system, hence the dimensionality of  $\mathbf{T}_x$  will be  $N \times 5$ . For the scope of this work, it is natural that only some of the components will later contribute to  $\mathbf{Y}$ ; those are the components that will be called ‘selective components’. The components that are not involved in the construction of  $\mathbf{Y}$  are called ‘irrelevant’. Here, we have chosen three (out of five) components to be ‘selective’ and the other two as ‘irrelevant’. The first ones will here be indicated as ‘selective components’ ( $K_{sel}$ ) and the others will be called ‘irrelevant components’ ( $K_{irr}$ ). Therefore, the  $\mathbf{T}_x$  is built as the concatenation of  $\mathbf{T}_{x_{sel}}$  and  $\mathbf{T}_{x_{irr}}$  scores, where  $\mathbf{T}_{x_{sel}}$  represents the ‘selective scores’ based on the ‘selective components’, and  $\mathbf{T}_{x_{irr}}$  represents the ‘irrelevant’ ones. These two matrices will have dimensions  $(N \times K_{sel})$  and  $(N \times K_{irr})$ , respectively. Consequently, the  $\mathbf{T}_x$ -matrix is built as:

$$\mathbf{T}_x = [\mathbf{T}_{x_{sel}} \mathbf{T}_{x_{irr}}] \quad (\text{A.1})$$

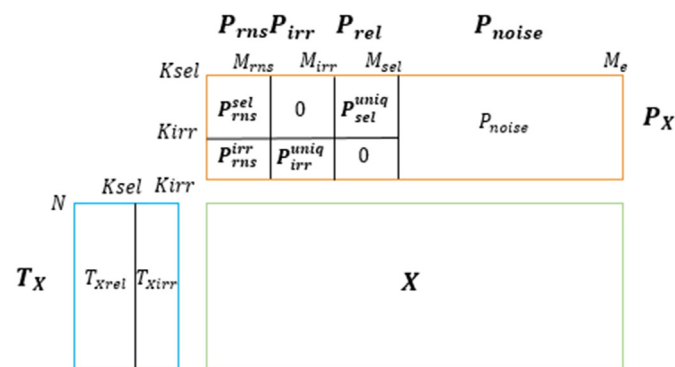
Then, the coefficient vector  $\mathbf{b}$  ( $K_{sel} \times 1$ ) is generated as a matrix containing random values drawn from the uniform distribution in the open interval (0.05, 1.05).

The response  $\mathbf{Y}$  is built as:

$$\mathbf{Y} = \mathbf{T}_{x_{sel}} \mathbf{b} \quad (\text{A.2})$$

Therefore, only the ‘selective’ scores are involved in the creation of  $\mathbf{Y}$ .

As for the scores, the distinction between a ‘selective’ and an ‘irrelevant’ part will apply also to the  $\mathbf{X}$ -loadings ( $\mathbf{P}_x$ ). In particular, in order to produce simulated data closer to real data, loadings will not only have a ‘selective’ and an ‘irrelevant’ part, but they will also have a part that is ‘relevant but not selective’ and some noise variables. The ‘relevant but not selective’ part is built by overlapping *selective* and *irrelevant* information, as shown above and in Fig. A.1. All the different types of variables that constitute the loadings are variables generated using the



**Fig. A.1.** Graphical representation of the simulation of the matrices  $\mathbf{X}$ ,  $\mathbf{T}_x$  and  $\mathbf{P}_x$ . The figure shows the partition of  $\mathbf{P}_x$  in ‘relevant but not selective’-variables  $\mathbf{P}_{rns}$ , ‘irrelevant’-variables  $\mathbf{P}_{irr}$ , ‘selective’-variables  $\mathbf{P}_{sel}$ , and noise-variables  $\mathbf{P}_{noise}$ ; and their specific dimensions.  $\mathbf{P}_{rns}$ ,  $\mathbf{P}_{irr}$  and  $\mathbf{P}_{sel}$  are partitioned matrices.  $\mathbf{P}_{rns}$  is constituted by the concatenation of  $\mathbf{A}_{sel}$  ( $K_{sel} \times M_{rns}$ ) and  $\mathbf{A}_{irr}$  ( $K_{irr} \times M_{rns}$ ).  $\mathbf{P}_{irr}$  is partitioned in a submatrix of zeros and  $\mathbf{B}_{irr}$  ( $K_{irr} \times M_{irr}$ ).  $\mathbf{P}_{sel}$  is partitioned in  $\mathbf{B}_{sel}$  ( $K_{sel} \times M_{sel}$ ) and a submatrix of zeros.  $\mathbf{T}_x$ -scores matrix is made by the concatenation of  $\mathbf{T}_{x_{sel}}$  ( $N \times K_{sel}$ ) and  $\mathbf{T}_{x_{irr}}$  ( $N \times K_{irr}$ ). More details on the submatrices can be found in the text.

normal distribution  $N(0,1)$ .

This means that each block will be constituted of a certain amount of ‘selective’-variables, ‘irrelevant’-variables, ‘relevant but not selective’-variables and some noise variables. More details about the structure of the loadings are reported below. The total number of variables is fixed for each block, but the relative amount of the different “type” of variables varies according to the design (the number of noisy variables is changing in order to sum up to the total). Here, we denote the number of ‘selective’ variables, ‘irrelevant’ variables, ‘relevant but not selective’ variables and the noise variables by call  $M_{sel}$ ,  $M_{irr}$ ,  $M_{rns}$  and  $M_e$ . The



'relevant but not selective'-loadings matrix of dimension ( $K \times Mrns$ ) is denoted by  $\mathbf{P}_{rms}$ , the 'selective'-loadings matrix of dimension ( $K_{sel} \times M_{sel}$ ) is denoted by  $\mathbf{P}_{sel}$ , the 'irrelevant'-loadings matrix of dimension ( $K_{irr} \times M_{irr}$ ) is denoted by  $\mathbf{P}_{irr}$  and the part representing the noise variables by  $\mathbf{P}_{Noise}$ .

The  $\mathbf{P}_{rms}$  is a block matrix of the form:

$$\mathbf{P}_{rms} = \begin{bmatrix} \mathbf{P}_{rms}^{sel} \\ \mathbf{P}_{rms}^{irr} \end{bmatrix} \quad (\text{A.3})$$

where  $\mathbf{P}_{rms}^{sel}$  ( $K_{sel} \times Mrns$ ) and  $\mathbf{P}_{rms}^{irr}$  ( $K_{irr} \times Mrns$ ) are matrices of random numbers normally generated. Performing the TP-product to create the  $\mathbf{X}$ -block, the sub-matrix  $\mathbf{P}_{rms}^{sel}$  will be multiplied by  $\mathbf{T}_{X_{sel}}$ , while  $\mathbf{P}_{rms}^{irr}$  is the part multiplied by  $\mathbf{T}_{X_{irr}}$ . This creates an overlapping between the 'selective' and the 'irrelevant' information. The  $\mathbf{P}_{irr}$  and  $\mathbf{P}_{sel}$  are in the form:

$$\mathbf{P}_{irr} = \begin{bmatrix} 0 \\ \mathbf{P}_{irr}^{uniq} \end{bmatrix} \text{ and } \mathbf{P}_{sel} = \begin{bmatrix} \mathbf{P}_{sel}^{uniq} \\ 0 \end{bmatrix} \quad (\text{A.4})$$

where  $\mathbf{P}_{irr}^{uniq}$  has dimensions ( $K_{irr} \times M_{irr}$ ) and  $\mathbf{P}_{irr}$  will be of dimensions: ( $K \times M_{irr}$ ).

$\mathbf{P}_{sel}^{uniq}$  has dimension ( $K_{sel} \times M_{sel}$ ) and  $\mathbf{P}_{sel}$  will be of dimensions ( $K \times M_{sel}$ ). The  $\mathbf{P}_{Noise}$  consist of zeros only.

This means that  $\mathbf{P}_X$  can then be represented as a partitioned matrix of the form:

$$\mathbf{P}_X^T = \begin{bmatrix} \mathbf{P}_{rms} & \mathbf{P}_{irr} & \mathbf{P}_{sel} & \mathbf{P}_{Noise} \\ (K \times Mrns) & (K \times M_{irr}) & (K \times M_{sel}) & (K \times Me) \end{bmatrix} \quad (\text{A.5})$$

Fig. A.1 gives a graphical illustration of how the loadings  $\mathbf{P}_X$  are partitioned.

Then, the  $\mathbf{X}$ -block can be calculated:

$$\mathbf{X} = \mathbf{T}_X \mathbf{P}_X^T \quad (\text{A.6})$$

Noise is added to the  $\mathbf{X}$ - and  $\mathbf{Y}$ -blocks. For  $\mathbf{Y}$ , the noise corresponds to a certain percentage of the standard deviation of  $\mathbf{Y}$  as reported below in Appendix B. For  $\mathbf{X}$ , the standard deviation for each column of  $\mathbf{X}$  is first calculated. Then, the pooled standard deviation is calculated, but only taking into account the columns that are not related to the noisy variables. In conclusion, the noise that is added to the  $\mathbf{X}$ -block is a certain percentage (according to the design), of this pooled standard deviation.

The test set for the external validation is built in the same way, but the number of samples ( $N_t$ ) is higher. The dimensionality of  $\mathbf{X}_t$  and  $\mathbf{Y}_t$  is fixed; these are  $1000 \times 400$  and  $1000 \times 1$ , respectively. The  $\mathbf{X}$ -scores for the test set  $\mathbf{T}_{X_{test}}$ , are generated as before and have dimensions ( $N_t \times K$ ).

The distinction among the variables that has been defined for the training set also applies to the test sets.  $\mathbf{Y}_t$  is calculated by the selective scores for the test set,  $\mathbf{T}_{X_{sel}}$ :

$$\mathbf{Y}_t = \mathbf{T}_{X_{sel}} * \mathbf{b} \quad (\text{A.7})$$

Since the loadings are the same as in the training set,  $\mathbf{X}_t$  is calculated as:

$$\mathbf{X}_t = \mathbf{T}_{X_{test}} \mathbf{P}_X^T \quad (\text{A.8})$$

and noise is added in the same way as above.

Dataset-2 is simulated in the same way as Dataset-1. The difference between the data sets is only in the dimensions. The number of rows of  $\mathbf{X}$  in Dataset-2 varies following the design described in Appendix B, while the number of columns is fixed to 40.

## Appendix B. Design of the experiment, methods and model parameters

### B.1. Experimental design for simulations

The experimental design for the study in Appendix A (for selecting the best variable selection methods) consists of seven factors with different numbers of levels. The seven factors are:

1. Variable selection method
2. Number of samples ( $N$ )
3. Number of 'relevant but not selective variables' ( $Mrns$ )
4. Number of 'selective variables' ( $M_{sel}$ )
5. Number of 'irrelevant variables' ( $M_{irr}$ )
6. Noise added to the  $\mathbf{Y}$  vector
7. Noise added to  $\mathbf{X}$ .

The factor 'Variable selection method' has eight levels. These are the PLS regression for the full model plus the following seven selection methods:

1. VIP
2. Selectivity Ratio
3. Jackknifing
4. sMC
5. UVE
6. Trunc-PLS
7. Forward Selection.

These variable selection methods can be mainly divided into methods based on the observation of model parameters and statistical/chemometric approaches. Below follows a brief description of each of them.

#### B.1.1. Variable selection methods based on the observation of the estimated model parameters

If a model is reliable, its parameters are good indicators of the sources of variation. Therefore, the regression coefficients and the loadings can be used to get indications of which variables are influencing the model strongly. When these estimated values are close to zero, the associated variables are presumably not relevant, at least together with all the other variables in the model. Estimated model parameters can also be used to calculate indicators that show which predictors are the more relevant (or less relevant).

**Selectivity Ratio (SR).** The so-called Selectivity Ratio (SR) [18] is the ratio between the variance explained by each predictor and the residual variance. The approach pursued in the present work, is the one proposed by Kvalheim in [20]. In the literature, there are different ways of defining cut-off values. In this work, two cut-off values will be used. One of them is the one proposed in [20] and it is based on a threshold calculated on the basis of an  $F$ -test (with fixed false-rejection probability at 0.05). For each variable, the corresponding Selectivity Ratio  $SR_j$  is defined as the ratio of two variances and, therefore, under the null hypothesis should be distributed as an  $F$ -distribution with  $N - 2$  and  $N - 3$  degrees of freedom, respectively [20]. Accordingly, if a  $SR_j$  is greater than the critical value of the  $F$ -distribution, the corresponding variable is considered significant and it is selected. Nevertheless, the application of a cut-off value based on the  $F$ -test is not always the most appropriate choice. For some data, this is a too parsimonious criterion. This is an issue recognized and discussed in [31].

Consequently, SR's mean is here proposed as an alternative cut-off value, to be used when this problem arises. In the present paper, this alternative cut-off value is used for the simulated multi-block data sets (Section 5.1). Both cut-off values have been used and compared for the flavored waters data set (Section 5.2). For the spectroscopy data set, the cut-off based on the  $F$ -test has been preferred. Also in this case both were used, but appeared that the cut-off based on the mean was influencing negatively the predictions.

### Variable Importance in Projection (VIP)

The *Variable Importance in Projection* (VIP) [15,17] is another model-based method widely used to select features. VIP is a measure of how much of the variance of  $\mathbf{X}$  is explained by each variable and, at the same time, of the  $\mathbf{X}$ 's correlation with  $\mathbf{Y}$ . The mean of the squared VIP scores, by construction, is equal to one. Variables with a VIP bigger than one are considered the most relevant (and therefore those are selected).

### Significance Multivariate Correlation (sMC)

Significance Multivariate Correlation (sMC) is a method that has been developed in order to estimate, for each variable, the sources of variability coming from a PLS-regression [21]. In order to assess which variables are important for the regression purpose, the ratios between the variable-wise Mean Squared Errors (MSE) of the PLS model and the mean squared of its residuals are compared to an  $F$ -test with 1 and  $N - 2$  degrees of freedom [21]. The variables that exceed the  $F$ -test threshold are selected.

### Elimination of Uninformative Variables for multivariate calibration (UVE)

The method is based on the analysis of the regression coefficients obtained from a PLS-regression of  $\mathbf{Y}$  on  $\mathbf{X}$  [22]. Those are then compared to the regression coefficients of a second regression, in which  $\mathbf{Y}$  is fitted to an  $\mathbf{X}\mathbf{R}$  matrix of dimensions  $N \times 2J$  (where the last  $J$  variables are generated randomly). Then, an entity called *reliability*  $c_j$  (based on regression coefficients) is defined [22]. The variables that will result in a *reliability* bigger (in absolute value) than random variables' reliability are selected.

**Truncation PLS.** Truncation-PLS can be based on different regression parameters. In this work it is based on loading weights, as suggested in [23]. The method is based on the idea that if a variable is uncorrelated to the response, loading weights will be equally distributed random variables, not different from random normal noise. Otherwise, they are normally distributed but with non-zero mean. Feature selection is conducted by observing which variables deviate from the median of the loading weights.

**Forward selection.** The forward selection approach starts with no variables in the model and then tests the inclusion of each variable by the means of a specific criterion [24]. The process is repeated until no variable improves the model. When the number of the variables is high, e.g. in spectroscopy, it is more reasonable, to perform the forward selection on intervals instead of on each variable.

**Jackknifing.** Jackknifing is a resampling procedure that can also be used for significance testing. The basic idea behind the method is that the uncertainty of a specific parameter is estimated by leaving out one observation at a time [25]. In this work, the estimated parameters are the regression coefficients. The uncertainty has been calculated following the modification to the original method by Martens et al. in [26].

Levels related to the other factors are reported in Table B.1 for both data sets.

**Table B.1**

Levels of six factors of the experimental design used (Factors: number of samples, number of relevant but non-selective variables, number of selective variables, number of irrelevant variables, noise added to the  $\mathbf{Y}$  vector, noise added to  $\mathbf{X}$ ) for both datasets. The missing factor in the table, the variable selection method, is illustrated in the text.

Dataset	# samples	# Relevant but non-selective variables	# Selective variables	# Irrelevant variables	Noise of $\mathbf{Y}$ (%)	Noise of $\mathbf{X}$ (%)
Dataset-1	10	10	10	10	15	10
	50	50	50	50	25	20
	100	100	100	100	35	30
Dataset-2	15	5	5	5	20	10
	30	10	10	10	30	20

At the end, following a full factorial design, 5832 ( $3^6 \times 8$ ) experiments are simulated for Dataset-1 and 512 ( $2^6 \times 8$ ) for Dataset-2.

### B.2. Evaluation criteria for assessing the PLS variable selection methods

Dataset-1 and Dataset-2 have been simulated following the above design repeated one hundred times. The ANOVA analysis that follows is based on the averages over these replicates. Following the full factorial design described above, PLS-regression models using all the variables were built and then the different selection methods have been applied. After the application of each variable selection method, a new PLS-regression using the selected variables has been performed. Different properties of the models were investigated. Many of these properties are expressed as relative percentages of a specific type of variables. This means that this value corresponds to the ratio between the number of a specific type of selected variables and the total number of that type of variables in the data set multiplied by 100. E.g., the relative percentage of 'selective' variables selected is calculated as the ratio between the number of the selected 'selective' variables and the total number of the 'selective' variables in the data set multiplied by 100. The same is done for the other types of variables.

The different properties investigated are:

- The explained test set variance of  $\mathbf{Y}$
- Relative percentage of 'selective' variables selected ( $R_{sel}$ )
- Relative percentage of 'irrelevant' variables selected ( $R_{irr}$ )
- Relative percentage of the 'relevant but not selective' variables selected ( $R_{rms}$ )
- Relative percentage of noise-variables selected ( $R_{noise}$ )
- Relative percentage of total variables selected ( $R_{tot}$ ).

#### B.2.1. ANOVA analysis

The ANOVA analysis performed included all the factors plus all the possible two-way interactions. Concerning Dataset-1, all the factors in the ANOVA are significant (independent of which property it was based on). This assumption is based on p-values, using a significance level of 5%. Concerning the interactions, all are significant, except interactions between 'selective' and 'relevant but not selective', 'irrelevant' and 'selective', and 'selective' and Noise  $\mathbf{X}$ .

Averaged RMSEPs,  $R_{sel}$ ,  $R_{irr}$ ,  $R_{rms}$  and  $R_{tot}$  for each variable selection method are reported in Table B.2. These values are grand means obtained by averaging across the (one hundred) replicates and the (729) models.

**Table B.2**

Dataset-1: Means (over all the experiments) of RMSEP,  $R_{rms}$ ,  $R_{irr}$ ,  $R_{sel}$ , and  $R_{tot}$  for each variable selection method.

	RMSEP	$R_{rms}$	$R_{irr}$	$R_{sel}$	$R_{tot}$
VIP	0.141	66	8	58	16
SR	0.141	60	0	82	18
JK	0.143	76	69	90	30
SMC	0.144	67	70	92	28
UVE	0.145	65	57	85	26
Trunc-PLS	0.144	59	8	57	14

PLS-models (both with or without variable selection) result in an averaged (grand mean across replicates and models) RMSEP of 0.14. Also the explained  $\mathbf{Y}$ -variance of PLS on the full models (all the variables are used) is comparable to the explained variance from models after the variable selection.

Investigating deeply data, it comes out that, when the noise in  $\mathbf{Y}$  is at the lower level (15% of the standard deviation of  $\mathbf{Y}$ ), the averaged (over the replicates) explained variances are 85% both for the full and the reduced models. This means that all the variance that could be modeled is actually captured by the models. Similarly, when the noise in  $\mathbf{Y}$  is at the highest level (35%), the averaged explained variance is 65%.

For the number and type of selected variables, the various variable selection methods show different behavior. All the methods select high percentages of 'relevant but not selective' variables which is an attractive property. The one that selects less variables is Trunc-PLS (59%), but the one that selects the most (Jackknifing) selects 76%, so the differences are not dramatic. Some methods, such as jackknifing, SMC and UVE select high percentages of total variables. Nevertheless, they present high percentages of selected variables of all types. Consequently, they are those that select more 'selective' variables but, at the same time, they select many 'irrelevant' ones (both systematic and noise). SR (and to a lesser extent), VIP and Trunc-PLS skip the systematic but 'irrelevant' variables which is an interesting property. These three methods are also the best in avoiding the selection of noise (VIP in particular). Hence, SR is in general the best at avoiding inclusion of unrelated information and maintaining the relevant ones.

In order to investigate whether the different variable selection methods behave differently at the different points of the design, also results averaged only over the one hundred replicates have been inspected (So, in this case they are averaged only over replicates and not over the 729 models). Consequently, specific trends for each variable selection method were pointed out. For instance, VIP is skipping less 'irrelevant' variables when the different types of variables ('selective', 'irrelevant' and 'relevant but not selective') are at the lowest levels. In these cases, it selects around 20% of the 'irrelevant' variables. This ability does not seem to be affected by the level of the noise in  $Y$ . Concerning the jackknifing, it seems to be more influenced by the level of the noise. It selects less 'irrelevant' variables (both systematic and noise) when the noise in  $X$  and in  $Y$  are at the lowest levels. The SMC method is not good at skipping the 'irrelevant' variables when there are few of them (lowest level) regardless of noise level in  $Y$ . The averaged amount of 'irrelevant' variables selected in these cases is 88%. UVE has good performance; it is particularly efficient in skipping a high percentage of 'irrelevant' variables when the number of the 'selective' and 'relevant but not selective' is high. In the same points, it selects also a high percentage of 'selective' and 'relevant but not selective' variables. Finally, t-PLS is not very stable in its selection, so it is not showing a clear trend.

Also in *Dataset-2*, all the factors are significant in the ANOVA analysis. Regarding the interactions, those between *method* and the other factors are all significant. Interactions between *number of samples* and the other factors are significant except for the interaction between *number of samples* and *Noise X* and the interaction between *number of samples* and *relevant variables*. All the other interactions are non-significant. Consequently, it appears that, reducing the dimensions of the data sets, the interactions between the different types of variables have no significant effect on the models (because all the possible interaction between *Rrns*, *Rirr* and *Rrel* are non-significant). This is an indication that, at these conditions, models are mainly dominated by factors *method* and *number of samples*.

Concerning the percentages of selected variables, the different methods follow trends similar to those presented for *Dataset-1*.

In conclusion, the methods show in general high ability in selecting relevant variables in the simulated data sets. Nevertheless, each of them has specific characteristic that would make it more suitable than other ones in different situations. For example, to avoid including information from non-related interferences, the best choice would be to use a method that is able to remove the systematic-'irrelevant' variables. Therefore, the choice would fall on SR, VIP and Trunc-PLS. On the other hand, if data are highly affected by non-systematic noise, the best option would be VIP, while the most unsuitable would be jackknifing.

As can be seen, there are many aspects that characterize a good method for variable selection, therefore, a compromise is required.

Ideally, from the interpretation point of view, the "best" method is the one that gives high values of *Rsel*, *Rrns* and low values of *Rirr* and *Rnoise*. For prediction purposes, the "best" method is the one giving a small RMSEP or a high explained variance.

Below, we will develop an approach based on a desirability index for a combined look at all the aspects.

## B.2.2. Selection of the most appropriate variable selection method

**B.2.2.1. Desirability index.** The desirability index (*di*) proposed here is based on the relative percentage of 'selective' variables (*Rsel*), relative percentage of 'irrelevant' variables (*Rirr*), relative percentage of the 'relevant but not selective' variables selected (*Rrns*) and relative percentage of Noise-variables selected (*Rnoise*). In this case, all of them are used as fractions between zero and one. This index is conceived to point out the "best" method from the interpretation point of view, therefore, explained variances or RMSEPs are not involved.

*Rsel* and *Rrns* were used as they are (since a high value of these is considered to have a good influence on the final model). For the 'Irrelevant' variables and the noise,  $1-Rirr$  and  $1-Rnoise$  were used to calculate the index.

The desirability index is calculated by taking the geometric average of those quantities in the 729 (for *Dataset-1*) and 64 (for *Dataset-2*) different points of the designs. The closer to 1 the index is, the better the method is performing.

Another desirability index is also calculated, focusing more on predictions and on removal of 'irrelevant' variables. This is done to check if developing the index from a more prediction-oriented prospective could give different results. Consequently, the additional index is based on averaged explained variance, *Rirr* and *Rnoise*. The two indices are in agreement, therefore, only results for *di* are shown and discussed.

*di*'s values for *Dataset-1* are reported in Table B.3. The highest values were obtained for SR and VIP (in decreasing order) which fits well with the observations from the ANOVA above. Consequently, these are the two chosen methods to be applied to the multi-block data sets. Concerning the other methods, Trunc-PLS gives a slightly lower value than VIP. Jackknifing's and UVE's values are comparable and a bit lower than Trunc-PLS'. This is due to the high amount of 'irrelevant' variables selected by these methods. Finally, SMC is the one giving the lowest *di*.

**Table B.3**

Desirability indices for each variable selection method: desirability indices are calculated as the geometric means of four properties (relative percentage of selective variables (*Rsel*), relative percentage of irrelevant variables (*Rirr*), relative percentage of the relevant but non-selective variables selected (*Rrns*) and relative percentage of noise-variables selected (*Rnoise*) for each variable selection method present in the design. The two highest desirability indices are reported in bold.

Method	VIP	SR	Jackknifing	sMC	tPLS	UVE
<i>di</i>	<b>0.77</b>	<b>0.84</b>	0.60	0.55	0.72	0.65

The desirability index was also calculated for *Dataset-2*; the same method appeared to be the most recommended. Therefore, VIP and SR are used in the multi-block part of this study.

## B.3. Conclusions on the simulation study and prospective for inclusion in a multi-block regression context

Apparently, VIP and SR are the most suitable methods under the considerations presented in the previous paragraphs. From the prediction point of view, they give comparable results. Considering the interpretation, the two methods reduce the amount of variables, but retain relevant ones. Both are powerful in skipping 'irrelevant' variables. In particular, SR is able to get rid of the systematic 'irrelevant' variables; which indicates that this method would be suitable to remove systematic errors in real data. The VIP is more efficient in removing random noise.

In addition to VIP and SR, also the forward selection method will be used for multi-block data sets. This is included in the work for the sake of completeness and to achieve a more general discussion. The forward

selection method will be used in two different versions: one selecting individual variables and one selecting windows of variables. The latter is suitable for highly collinear spectral data with very many variables.

## References

- [1] R. Bro, F. van den Berg, A. Thybo, C.M. Andersen, B.M. Jørgensen, H. Andersen, Multivariate data analysis as a tool in advanced quality monitoring in the food production chain, *Trends Food Sci. Technol.* 13 (2002) 235–244.
- [2] J. Pagès, Multiple factor analysis: main features and application to sensory data, *Rev. Colomb. Estad.* 27 (2004) 1–26.
- [3] S. Hassani, H. Martens, E.M. Qannari, M. Hanafi, G.I. Borge, A. Kohl, Analysis of -omics data: graphical interpretation- and validation tools in multi-block methods, *Chemom. Intell. Lab. Syst.* 104 (2010) 140–153.
- [4] S. Wold, S. Hellberg, T. Lundstedt, M. Sjöström, H. Wold, *Proc. Symp. on PLS Model Building: Theory and Application*, Frankfurt am Main, 1987; Also Tech. rep., Department of Organic Chemistry, Umea University, 1987.
- [5] J.C. Gower, Generalized Procrustes analysis, *Psychometrika* 40 (1975) 33–51.
- [6] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, *Chemom. Intell. Lab. Syst.* 124 (2013) 32–42.
- [7] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, *Food Qual. Prefer.* 24 (2012) 8–16.
- [8] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, *J. Chemom.* 25 (2011) 441–455.
- [9] E. Acar, E.E. Papalexakis, G. Gürdeniz, M.A. Rasmussen, A.J. Lawaetz, M. Nilsson, R. Bro, Structure-revealing data fusion, *BMC Bioinf.* 15 (2014) 239.
- [10] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, *J. Chemom.* 24 (2010) 728–737.
- [11] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, *Chemom. Intell. Lab. Syst.* 118 (2012) 62–69.
- [12] J.A. Westerius, T. Kourti, J.F. MacGregor, Analysis of hierarchical PCA and PLS models, *J. Chemom.* 12 (1998) 301–321.
- [13] T. Næs, O. Tomic, B.-H. Mevik, H. Martens, Path modelling by sequential PLS regression, *J. Chemom.* 25 (2011) 28–40.
- [14] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, *Chemom. Intell. Lab. Syst.* 141 (2015) 58–67.
- [15] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, *3D QSAR in Drug Design*, 1 1993, pp. 523–550.
- [16] I.G. Chong, C.H. Jun, Performance of some variable selection methods when multicollinearity is present, *Chemom. Intell. Lab. Syst.* 78 (2005) 103–112.
- [17] S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, *Chemom. Intell. Lab. Syst.* 129 (2013) 76–86.
- [18] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, *Chemom. Intell. Lab. Syst.* 95 (2009) 35–48.
- [19] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multi-block PLS and PC models for easier model interpretation and as an alternative to variable selection, *J. Chemom.* 10 (1996) 463–482.
- [20] O.M. Kvalheim, Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots, 24 (2010) 496–504.
- [21] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC), *Chemom. Intell. Lab. Syst.* 138 (2014) 153–160.
- [22] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, *Anal. Chem.* 68 (1996) 3851–3858.
- [23] K.H. Liland, M. Høy, H. Martens, S. Sæbø, Distribution based truncation for variable selection in subspace methods for multivariate regression, *Chemom. Intell. Lab. Syst.* 122 (2013) 103–111.
- [24] N.R. Draper, H. Smith, *Applied Regression Analysis*, Wiley-Interscience, Hoboken, NJ, 1998 307–312.
- [25] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982 (ISBN 0-89871-179-7).
- [26] H. Martens, M. Martens, Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression, *Food Qual. Prefer.* 11 (2000) 5–16.
- [27] U. Indahl, T. Næs, Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling, *J. Chemom.* 12 (1998) 261–278.
- [28] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval Partial Least-Squares Regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, *Appl. Spectrosc.* (2000) 413–419.
- [29] N.K. Afseth, V.H. Segtnan, B.J. Marquardt, J.P. Wold, Raman and near-infrared spectroscopy for quantification of fat composition in a complex food model system, *Appl. Spectrosc.* 59 (2005) 1324–1332.
- [30] R.D. Snee, Validation of regression models: methods and examples, *Technometrics* 19 (1977) 415–428.
- [31] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles, *Anal. Chem.* (2009) 2581–2590.

## **Paper III**

**Title:** Extension of SO-PLS to multi-way arrays

**Authors:** A. Biancolillo, T. Næs, R. Bro, I. Måge

*Submitted to Chemometrics and Intelligent Laboratory Systems on June 29<sup>th</sup> 2016*



# Extension of SO-PLS to multi-way arrays: *SO-N-PLS*

Alessandra Biancolillo<sup>a,b,\*</sup>, Tormod Næs<sup>a,b</sup>, Rasmus Bro<sup>b</sup>, Ingrid Måge<sup>a</sup>

<sup>a</sup> *Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway*

<sup>b</sup> *Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

\*Corresponding author:

Tel: +47 64 97 01 15

e-mail: [alessandra.biancolillo@nofima.no](mailto:alessandra.biancolillo@nofima.no)

## Abstract

Multi-way data arrays are becoming more common in several fields of science. For instance, analytical instruments can sometimes collect signals at different modes simultaneously, as e.g. fluorescence and LC/GC-MS. Higher order data can also arise from sensory science, where product scores can be reported as function of sample, judge and attribute. Another example is process monitoring, where several process variables can be measured over time for several batches. In addition, so-called *multi-block data sets* where several blocks of data explain the same set of samples are becoming more common. The same samples are for instance analyzed by different techniques, or in different times and places. Several methods exist for analyzing either multi-way or multi-block data, but there has been little attention on methods that combine these two data properties. A common procedure is to “unfold” multi-way arrays in order to obtain two-way data tables on which classical multi-block methods can be applied. However, it is a known fact that unfolding can lead to overfitted models due to increased flexibility in parameter estimation. In this paper we present a novel multi-block regression method that can handle multi-way data blocks. This method is a combination of a multi-block method called Sequential and Orthogonalized-PLS (SO-PLS) and the multi-way version of PLS, *N-PLS*. The new method is therefore called *SO-N-PLS*. We have compared the method to Multi-block-PLS (MB-PLS) and SO-PLS on unfolded data. We investigate the hypotheses that *SO-N-PLS* has better performances on small data sets and noisy data, and that *SO-N-PLS* models are easier to interpret. The hypotheses are investigated by a simulation study and two real data examples; one dealing with regression and one with classification. The simulation study shows that *SO-N-PLS* performs better than the unfolded methods when the sample size is small and the data is noisy. This is due to the fact that it filters out the noise better than MB-PLS and SO-PLS. For the two real data examples, the superiority of *SO-N-PLS* method is not so noticeable, but it performed well. Additionally, *SO-N-PLS* gives rise to a number of graphical interpretation tools, which are described and discussed in the paper.

## Keywords

Multi-block; *N-PLS*; SO-PLS; multi-way; MB-PLS; regression; classification

## 1. Introduction

Data tables with more than two modes are called *multi-way arrays*, and can arise from many different fields in modern science. Important examples of three-way arrays (the most common multi-way arrays) are data from various analytical instrumental techniques, e.g. spectroscopy (NMR, EEM), chromatography (GC/LC-MS) and multispectral imaging. In addition, time series data from, for instance, process monitoring (modes: batch, variable, time) and environmental analysis (modes: location, variable, time) are three-way. Sensory data can also be collected in a three-way array (modes: samples, judges and attributes), and data from experimental designs can be reported as functions of the experimental factors [1-3] with the factors representing the different ways.

The most common approach for handling multi-way arrays is to reorganize data in a two-way array. This is called *unfolding*, and it can be performed in different ways. In the case of three-way arrays we have *row-wise unfolding*, *column-wise unfolding* and *tube-wise unfolding*. Applying the first approach, a three-way array  $\underline{X}$  of dimensions  $N \times J \times K$  can be unfolded to a matrix  $\mathbf{X}$  of dimensions  $N \times (JK)$ . In the *column-wise* approach, a three-way array  $\underline{X}$  of dimensions  $N \times J \times K$  is unfolded to a matrix  $\mathbf{X}$  of dimensions  $(NK) \times J$ . Finally, in the *tube-wise* approach, the  $\mathbf{X}$  matrix's dimensions become  $(NJ) \times K$ .

The unfolding procedure makes multi-way arrays suitable for classical multivariate data analysis, but there are also some drawbacks with this approach. Firstly, model building using unfolded matrices can lead to overfitting since the number of estimated model parameters increases, often without improving the predictive power. Hence, the increased complexity is mainly used for fitting noise. Interpretation can also be more difficult when the original data structure is lost, both because of overfitting and of the increased number of parameters. Multi-way methods have been developed to overcome these drawbacks. PARAFAC [4-5], *N*-PLS [6] and Tucker-2 and Tucker-3 models [7] are some of the main methods that retain the original dimensions of a multi-way array.

It is often relevant to join multiple blocks of data in order to understand and exploit all the actual information on the system at study. *Multi-block methods* can handle several blocks of data at the same time, and examples of such methods are Multi-Block-PLS (MB-PLS), Sequential and Orthogonalized Partial Least Squares (SO-PLS), Parallel Orthogonalized Partial Least Squares (PO-PLS), OnPLS and Coupled Matrix and Tensor Factorization (CMTF) [8-14]. This is a new and emerging field, and many research challenges remain unsolved.

Previous work [11,15] has shown that the SO-PLS regression method provides similar (and sometimes better) predictions than MB-PLS, and that it has some properties that makes it particularly useful for interpretation purposes. So far, SO-PLS has been developed for two-way arrays only. In this paper we show how SO-PLS and *N*-PLS for multi-way regression can be combined to form a new regression method that we call SO-*N*-PLS. It can be used to analyze multiple multi-way predictor blocks or a combination of multi-way and two-way blocks. In Fig. 1 we show a graphical representation of the data structures that can be handled by SO-*N*-PLS, and also how they can be unfolded in order to be analyzed by two-way methods. In this paper the focus will be on two- and three-way arrays, but more general situations can also be handled within the same framework.

We will discuss how SO-*N*-PLS can be applied to both regression and classification problems. The novel method will be compared to SO-PLS and MB-PLS on unfolded data, and we will in particular investigate the following hypotheses:

- SO-N-PLS can provide better predictions than unfolded analysis. This is especially so for small sample sizes and noisy data, since the risk of overfitting is higher.
- SO-N-PLS can give simpler or improved interpretation than unfolded analysis.

In order to investigate these aspects, both real data and a simulation study will be used.

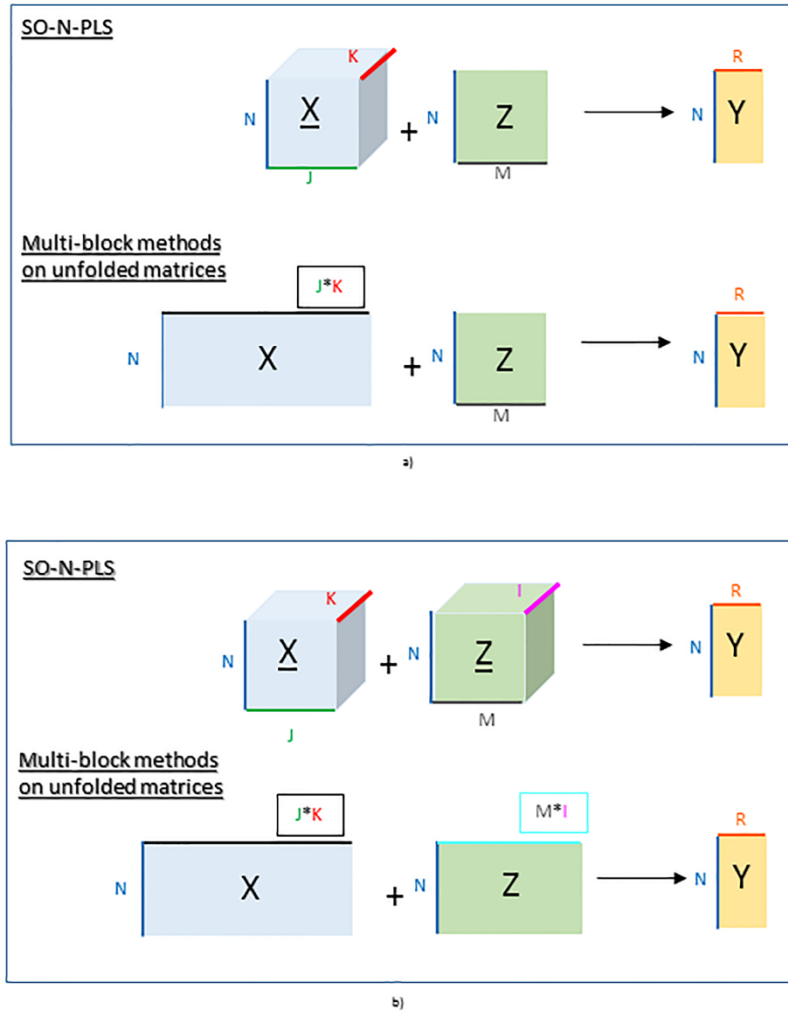


Figure 1: Graphical representation of SO-N-PLS and the multi-block methods on unfolded data. SO-N-PLS can be directly applied on the multi-way arrays avoiding unfolding. a) the  $\mathbf{X}$ -block is three-way, while  $\mathbf{Z}$  is a matrix. b) both  $\mathbf{X}$ - and  $\mathbf{Z}$ -blocks are three-way arrays. SO-PLS and MB-PLS are always applied on the row-wise unfolded matrices.

## 2. Material and methods

### 2.1 Sequential and Orthogonalized Partial Least Squares (SO-PLS) regression

Sequential and Orthogonalized Partial Least Squares (SO-PLS) [11] is a multi-block regression method for multiple predictor blocks. The model is assumed to be linear, and with two blocks the general formula is:

$$\mathbf{Y} = \mathbf{X}\mathbf{g} + \mathbf{Z}\mathbf{h} + \mathbf{E} \quad (1)$$

where  $\mathbf{X}_{(N \times J)}$  and  $\mathbf{Z}_{(N \times M)}$  are the two predictors blocks,  $\mathbf{Y}_{(N \times R)}$  is the response (categorical if the model is used for classification),  $\mathbf{g}_{(J \times R)}$  and  $\mathbf{h}_{(M \times R)}$  are the regression coefficients for each



of the two blocks and  $\mathbf{E}_{(N \times R)}$  is the residual matrix. In all cases, the data sets are assumed to be centered.

In this work we consider multi-block models with two predictor blocks, but it is straightforward to extend the method to more than two blocks [11].

The SO-PLS algorithm with two predictor blocks requires four steps (in addition to centering and possibly scaling the data):

1.  $\mathbf{Y}$  is fitted to  $\mathbf{X}$  by PLS-regression, giving PLS scores  $\mathbf{T}_X$ .
2.  $\mathbf{Z}$  is orthogonalized with respect to the scores  $\mathbf{T}_X$  from step 1 (obtaining  $\mathbf{Z}_{\text{orth}}$ ):
$$\mathbf{Z}_{\text{orth}} = \mathbf{Z} - \mathbf{T}_X(\mathbf{T}_X^T \mathbf{T}_X)^{-1} \mathbf{T}_X^T \mathbf{Z} \quad (2)$$
3. Residuals from the first PLS are fitted to  $\mathbf{Z}_{\text{orth}}$ , giving PLS scores  $\mathbf{T}_{Z_{\text{orth}}}$ .
4. The full predictive model is computed as the ordinary least squares fit of  $\mathbf{Y}$  to  $\mathbf{T}_X$  and  $\mathbf{T}_{Z_{\text{orth}}}$ . Since the first set of scores are linear functions of  $\mathbf{X}$  and the second set of scores are linear functions of  $\mathbf{Z}_{\text{orth}}$  which again is a linear function of  $\mathbf{Z}$ , this means that the equation can be reformulated into Eq. (1).

When more blocks are available, the procedure can be repeated as explained in [11]. Compared to MB-PLS, SO-PLS has the benefit of being invariant to block scaling, it allows different numbers of components from each block, and it permits individual interpretation of the contributions of each block. The  $\mathbf{X}$ -block is interpreted by looking at the first PLS model. The  $\mathbf{Z}$ -block can be interpreted by looking at the scores  $\mathbf{T}_{Z_{\text{orth}}}$  obtained in step 3 and the loadings obtained by regressing  $\mathbf{Z}$  onto  $\mathbf{T}_{Z_{\text{orth}}}$ .

SO-PLS can be combined with Linear Discriminant Analysis (LDA) [16] in order to create classification models [15]. The method is then called SO-PLS-LDA, and the only difference is that LDA is applied to the concatenated scores  $[\mathbf{T}_X \mathbf{T}_{Z_{\text{orth}}}]$  instead of ordinary least squares in step 4.

## 2.2 Multi-Block Partial Least Squares (MB-PLS) regression

Multi-block PLS is a well-established regression method [8-9]. The prediction model is estimated by classical PLS regression on the concatenated predictor blocks  $\mathbf{X}$  and  $\mathbf{Z}$ . In order to avoid that blocks of high dimensionality or large values dominate the model, data are usually scaled by dividing each block by its Frobenius norm. The PLS scores are called *super-scores*, and it is possible to calculate so-called *block-scores*, *block-weights* and *block-loadings* for interpretation purposes [8,9]. In the same way as classical PLS, MB-PLS can be used as a starting point for classification models. In this work, classification is performed by applying LDA to the super-scores [17]; we are going to refer to this method as MB-PLS-LDA.

## 2.3 N-PLS

N-PLS is a direct extension of classical PLS for multi-way arrays. In the three-way case it is called tri-PLS, and the bilinear decomposition of the predictor array is replaced by a tri-linear decomposition. For  $\underline{\mathbf{X}}_{(N \times J \times K)}$ , the  $F$ -component model corresponds to:

$$x_{njk} = \sum_{f=1}^F t_{nf} w_{jf}^J w_{kf}^K + e_{njk} \quad (3)$$

where  $\mathbf{t}$  are the scores and  $\mathbf{w}^J$  and  $\mathbf{w}^K$  are the weights of the second and third mode, respectively. The model corresponds to the so-called Martens PLS algorithm [18], in that there are no additional sets of loading vectors  $\mathbf{p}$  as in the two-way PLS algorithm. The loadings  $\mathbf{p}$  are

not used in N-PLS, as they would not provide orthogonality of the scores in the same way as in two-way PLS. The components are determined sequentially, and for each one the loading weights  $\mathbf{w}$  are found for the two variable modes in such a way that they provide scores  $\mathbf{t}$  that have maximum covariance with the still unexplained part of the response  $\mathbf{Y}$ .

The method can easily be extended to higher order data, and it can also be applied for more than one response variable; in which case it becomes iterative [6]. N-PLS can be combined with LDA for classification purposes. Similarly to MB-PLS-LDA, LDA is applied to the scores. We refer to this method as N-PLS-LDA.

## 2.4 SO-N-PLS

The algorithm proposed here combines the SO-PLS algorithm with N-PLS regression in order to build multi-block models with multi-way arrays as predictors. It is here presented only for three-way arrays, but as for N-PLS itself, it can easily be extended to arrays of a higher order.

The algorithm is the same as explained in paragraph 2.1, with the main difference that regressions which involve multi-way blocks are performed applying N-PLS instead of PLS. One then ends up with two sets of scores ( $\mathbf{T}_X$  and  $\mathbf{T}_{Z_{orth}}$ ) as for SO-PLS, and can run an ordinary least squares fit of  $\mathbf{Y}$  onto the scores. Note that it does not matter whether the three-way array is first or last. All the properties described for SO-PLS in section 2.1 are retained.

The orthogonalization in SO-N-PLS is slightly different than in SO-PLS when the second block is multi-way. The three-way  $\underline{\mathbf{Z}}$ -block of dimension  $N \times M \times I$  is first unfolded row-wise to  $\mathbf{Z}_{un}$ , a matrix of dimensions  $N \times MI$ . Then,  $\mathbf{Z}_{orth}$  is obtained replacing  $\mathbf{Z}$  with  $\mathbf{Z}_{un}$  in Eq.2 before  $\mathbf{Z}_{orth}$  is refolded back to the original three-way structure.

In order to obtain a regression equation in the original variables (instead of score vectors), the model needs to be formulated in terms of unfolded matrices. With two sets of unfolded matrices, i.e. if two three-way blocks are involved, the equation corresponding to Eq. 1 becomes:

$$\mathbf{Y} = \mathbf{X}_{un}\boldsymbol{\gamma} + \mathbf{Z}_{un}\mathbf{v} + \mathbf{E} \quad (4)$$

Where  $\mathbf{X}_{un}$ , is the unfolded  $\underline{\mathbf{X}}$ , a matrix of dimensions  $N \times JK$ .  $\boldsymbol{\gamma}_{(JK \times R)}$  and  $\mathbf{v}_{(MI \times R)}$  are the regression coefficients. Note that N-PLS involves two sets of weights,  $\mathbf{w}^J$  and  $\mathbf{w}^K$ . These weights are different from the weights extracted by PLS on an unfolded three-way matrix. Likewise, the  $\mathbf{g}$  and  $\mathbf{h}$  (from Eq.1) and  $\boldsymbol{\gamma}$  and  $\mathbf{v}$ , are not the same. Regression coefficients from SO-PLS and SO-N-PLS models have same size, but they are calculated differently.

Here, regression coefficient are calculated as suggested by De Jong in [19] (procedure called *Method 2*).

Firstly, weights  $\mathbf{W}$  are calculated as the inner product of the weights  $\mathbf{w}^J$  and  $\mathbf{w}^K$ . Then, the loading-weights  $\mathbf{R}$  are obtained as:

$$\mathbf{R} = \mathbf{W}/\delta \quad (5)$$

Where  $\delta$  is the upper triangular part of  $\mathbf{W}^T \mathbf{W}$ .

Then,  $\mathbf{b}_{N-PLS}$  can be calculated as:

$$\mathbf{b}_{N-PLS} = \mathbf{R} \mathbf{Q}^T \quad (6)$$

where  $\mathbf{Q}$  are the  $\mathbf{Y}$ -loadings.

SO-N-PLS can be used for classification problems by applying LDA on the concatenated  $\mathbf{T}_X$  and  $\mathbf{T}_{Z_{orth}}$ , as described for MB-PLS and SO-PLS. We call this method SO-N-PLS-LDA.

## 2.5 Estimating the number of optimal components in multi-block models

The number of latent variables to be used in each PLS regression can be decided by either a *global* or *sequential* strategy. In the global strategy, all combinations of components from each block are tested and evaluated using the so-called Måge-plot [11]. In the sequential strategy (not used here), the number of components to use for the first block is determined before the number of components for second block is assessed. With this strategy one extracts all relevant information from  $\mathbf{X}$  before  $\mathbf{Z}$  is introduced. In both cases, it is important to validate the model carefully since many combinations of components are tested.

In this work, the root mean square error of cross-validation (RMSECV) is used for selecting components in the regression models. For classification problems, the cross-validated classification error is used [15].

## 2.6 Graphical inspection of the model parameters

The interpretation tools used for SO-PLS are also applicable for SO-N-PLS, as for instance the interpretation of the scores plot discussed in [11,15]. The scores can be used to investigate the distribution of samples and look for clusters and groupings, just like for regular PLS. Scores can be plotted internally for each block, or scores from  $\mathbf{X}$  and  $\mathbf{Z}$  may be plotted against each other since they are all orthogonal.

As explained in Paragraph 2.3, N-PLS follows the Martens PLS algorithm, in which the weights  $\mathbf{W}$  are used to calculate the scores. For the  $\mathbf{X}$ -block, these weights can be used directly to interpret the variable contributions for each component in SO-N-PLS. For three-way arrays, there are two possible visualizations of the weights. One is obtained by plotting  $\mathbf{w}^J$  and  $\mathbf{w}^K$  individually. Another alternative is to plot the outer product of  $(\mathbf{w}^J \mathbf{w}^K)^T$  as a landscape.

For MB-PLS and SO-PLS,  $\mathbf{X}$ -loading weights for each component will be of size  $\mathbf{JK} \times \mathbf{1}$ . They can be plotted as they are, or folded back to an  $\mathbf{J} \times \mathbf{K}$  matrix and plotted as a landscape similarly to the outer product  $\mathbf{w}^J$  and  $\mathbf{w}^K$  for SO-N-PLS.

Interpretation of the  $\mathbf{Z}$ -block is slightly different than for the  $\mathbf{X}$ -block since  $\mathbf{Z}_{orth}$  is not in the row space spanned by  $\mathbf{Z}$ . In SO-PLS it has been shown that the  $\mathbf{Z}$ -block can be interpreted by calculating loadings  $\mathbf{P}_Z$  as projections of  $\mathbf{Z}$  itself on  $\mathbf{T}_{Z_{orth}}$  [15]:

$$\mathbf{P}_Z = (\mathbf{T}_{Z_{orth}}^T \mathbf{T}_{Z_{orth}})^{-1} \mathbf{T}_{Z_{orth}}^T \mathbf{Z} \quad (7)$$

In this way, loadings are showing the relation between  $\mathbf{Z}$  and the extracted information (after  $\mathbf{X}$  has been modelled).

In the three-way case, the  $\mathbf{Z}$ -weights can be re-calculated in a similar way by projecting the unfolded  $\mathbf{Z}$  on  $\mathbf{T}_{\text{Zorth}}$ :

$$\mathbf{W}_z = (\mathbf{T}_{\text{Zorth}}^T \mathbf{T}_{\text{Zorth}})^{-1} \mathbf{T}_{\text{Zorth}}^T \mathbf{Z}_{\text{un}} \quad (8)$$

By Eq. 8 we obtain unfolded  $\mathbf{W}_z$ . These can then be reshaped and plotted in the same ways as for  $\mathbf{W}_x$ .

Additionally, regression coefficients can be used to interpret variable contributions. One can, for instance, plot (one block at a time) regression coefficients from SO-N-PLS ( $\mathbf{y}$  and  $\mathbf{v}$  in Eq.4) and SO-PLS ( $\mathbf{g}$  and  $\mathbf{h}$  in Eq.1) as shown in Fig. 6. As for the weights, coefficients can be reshaped and plotted as landscapes.

## 2.7. Data analysis

All data analyses were performed using MATLAB (R2012b, The Mathworks, Natick, MA), using in-house routines. MATLAB routines for MB-PLS, SO-PLS, SO-N-PLS are available for download at [www.nofimamodeling.org](http://www.nofimamodeling.org).

## 3. Data sets

### 3.1 Simulated Data

Data sets consisting of two three-way predictor blocks ( $\mathbf{X}$  and  $\mathbf{Z}$ ) and a response vector ( $\mathbf{y}$ ) were simulated to investigate differences between SO-N-PLS, MB-PLS and SO-PLS under various scenarios. The data sets are constructed in such a way that they fit a low-dimensional three-way structure. The main focus is to compare the method performances on small and noisy data sets, since these are most prone to overfitting. Data sets were simulated following a full factorial design of the factors “*number of samples*” (six levels), and “*amount of random noise*” (four levels) ending up with  $6 \times 4 = 24$  different factor combinations. Noise was added to each variable of  $\mathbf{X}$  and  $\mathbf{Z}$ . The six different samples sizes ( $N_i$ ) are 15, 20, 25, 35, 50 and 60, while the four levels of added noise ( $L_1, L_2, L_3, L_4$ ) correspond to 10%, 30%, 40% and 50% of the signal. Noise was added also to  $\mathbf{y}$ , at a fixed level of 1.5% of the signal. All noise added was homoscedastic independent Gaussian.

Each factor combination was replicated one hundred times, resulting in  $6 \times 4 \times 100 = 2400$  different data sets. For each data set, an independent test set ( $\mathbf{X}_t, \mathbf{Z}_t$  and  $\mathbf{y}_t$ ) of 600 samples was constructed for validation purposes.

The three-way  $\mathbf{X}$ ,  $\mathbf{Z}$ ,  $\mathbf{X}_t$  and  $\mathbf{Z}_t$  predictor blocks were simulated to mimic fluorescence spectra, and were created in the following way:

$\mathbf{X}$  ( $N_i \times 201 \times 61$ ) is generated as the outer product of  $\mathbf{T}_x$ ,  $\mathbf{B}_x$  and  $\mathbf{C}_x$  while  $\mathbf{Z}$  ( $N_i \times 201 \times 61$ ) as the outer product of  $\mathbf{T}_z$ ,  $\mathbf{B}_z$  and  $\mathbf{C}_z$ . Scores  $\mathbf{T}_x$  and  $\mathbf{T}_z$  are both ( $N_i \times 2$ ) matrices of normally distributed random numbers.  $\mathbf{B}_x$  and  $\mathbf{B}_z$  (both  $201 \times 2$ ), and  $\mathbf{C}_x$  and  $\mathbf{C}_z$  (both  $61 \times 2$ ) are loadings extracted from real fluorescence spectra of mixtures of aminoacids (data set described in [20]). Consequently, the loading vectors are not orthogonal. Correlations between components within each loading are -0.21, -0.48, 0.93 and -0.15, for  $\mathbf{B}_x, \mathbf{B}_z, \mathbf{C}_x$  and  $\mathbf{C}_z$ , respectively.

The response vector  $\mathbf{y}$  is built as:

$$\mathbf{y} = [\mathbf{T}_X \mathbf{T}_Z] * \boldsymbol{\beta} \quad (9)$$

Where  $\boldsymbol{\beta}$  ( $4 \times 1$ ) is the coefficient vector generated as a matrix containing random values drawn from the uniform distribution on the interval (0.05, 1.05).

$\mathbf{T}_X$ ,  $\mathbf{T}_Z$  and  $\boldsymbol{\beta}$  (and consequently all the blocks) as well as the added noise are regenerated in each simulation.

### 3.2 Chemical mixture data set

28 samples of mixtures of five different biochemical compounds were analyzed by EEM and NMR. These compounds are two peptides, Valine-Tyrosine-Valine (*Val-Tyr-Val*) and Tryptophan-Glycine (*Trp-Gly*), a single amino acid, Phenylalanine (*Phe*), a sugar, Maltoheptaose (*Malto*), and an alcohol, *Propanol*. More details can be found in [21]. The two cubes of measures are used as  $\underline{\mathbf{X}}$  ( $28 \times 251 \times 21$ ) and  $\underline{\mathbf{Z}}$  ( $28 \times 13324 \times 8$ ) blocks in an SO-N-PLS regression model, in order to predict the concentration of compounds in the mixture. The same predictor blocks will be used to predict the five different responses  $\mathbf{y}_{TV}$ ,  $\mathbf{y}_{TG}$ ,  $\mathbf{y}_{Phe}$ ,  $\mathbf{y}_{Mal}$  and  $\mathbf{y}_{Pro}$  using five individual regression models. The response vectors correspond to the concentrations of *Val-Tyr-Val*, *Trp-Gly*, Phenylalanine, Maltoheptaose and Propanol, respectively.

### 3.3 Lambrusco data set

Lambrusco is a typical wine of the district of Modena (Italy) with protected denomination of origin (PDO). Lambrusco can be produced using mixtures of different species of grapes harvested in the area close to Modena. The fraction of the different grapes used is strictly fixed by the law under the PDO legislation. Unfortunately, frauds attempts in the food sector are quite common and wine is one of the main targets. Typical wine frauds can for instance be to use different fractions or lower quality grapes in PDO wines. Characterization and authentication of the grape cultivars used in wine production is therefore an important task, although not straightforward. In this work, the ability to distinguish between three different types of PDO Lambrusco wines based on instrumental analysis is tested. A total of fifty-eight samples of wines (all produced in 2009) were analyzed by EEM and NMR. Of these, nineteen are of "*Lambrusco Grasparossa di Castelvetro PDO*", twenty of "*Lambrusco Salamino di Santa Croce PDO*", and nineteen of "*Lambrusco di Sorbara PDO*". In the following analysis, the EEM three-way array is used as  $\underline{\mathbf{X}}$  ( $58 \times 161 \times 21$ ) while the NMR is used as  $\underline{\mathbf{Z}}$  ( $58 \times 9168$ ). SO-N-PLS-LDA model is then built to classify wines belonging to the three classes *Grasparossa*, *Sorbara* and *Salamino*. The response block is a categorical matrix carrying the class-belonging information. For a detailed description of the data set, see [22].

## 4. Results and Discussion

The simulated data sets were validated by independent test sets. The *mixture* and *Lambrusco* data sets were not considered large enough for a test set validation. Therefore, these models are validated by leave one out cross validation.

### 4.1 Results for the simulation study

SO-PLS and MB-PLS on the unfolded arrays and SO-N-PLS on the original data were performed on all the 2400 simulated data sets. The simulation study is divided in two parts: *Part I* and *Part II*, differing in how the model complexity is estimated. In the first one, the true number of latent

variables (LVs) is used, namely two for each block in SO(-N)-PLS and four in MB-PLS. This is done in order to compare the performances of models when the definition of optimal model complexity is not affecting the results. In part II, the numbers of components are selected for both blocks simultaneously using the Måge-plot (as described in paragraph 2.5). Instead of selecting the number of component resulting in the lowest RMSECV, an adjustment to ensure parsimony in the selection was carried out. The selected number/combination of components in MB-PLS/SO(-N)-PLS models is the smallest one giving an RMSECV not significantly different from the absolute minimum, decided by a  $\chi^2$  test (significance level 5%) [23]. *Part II* is more relevant for a real data analysis when the true complexity is unknown.

ANOVA was used to evaluate the effects of *Method*, *Samples* (N) and *Noise* (L) on the RMSEPs (averaged over 100 replicates). For results, see Table 1. In the simulation study Part I, the largest effects (as measured by MS's and F-values) are given by *Method* and *Samples*. For the simulation study Part II, *Samples* gives the largest effect followed by *Method* and *Noise*. It is clear that, relative to the number of samples, the effect of *Method* is smaller when selecting the number of components rather than knowing the 'correct' number a priori. This means that, when each method is allowed to find the optimal number of components, the differences between methods become smaller. Even though the underlying complexity of the two blocks is two, a different number of components could be optimal for the model. This aspect will be discussed further below.

Table 1: ANOVA analysis of RMSEP for the simulation studies.

		Simulation Part I			Simulation Part II		
Effect	D.o.f.	Mean Sq. ( $\times 10^{-3}$ )	F-value	p-value	Mean Sq. ( $\times 10^{-3}$ )	F-value	p-value
Method	2	3.1	21.0	0.000	0.8	6.3	0.005
Samples (N)	5	3.3	22.7	0.000	2.38	18.8	0.000
Noise (L)	3	0.5	3.2	0.037	0.65	5.2	0.005
Method*Samples	10	0.3	2.2	0.048	0.16	1.3	0.294
Method*Noise	6	0.1	0.6	0.707	0.08	0.6	0.721
Samples*Noise	15	1.3	9.1	0.000	1.06	8.4	0.00
Error	30	0.2			0.13		
R-squared	0.92					0.9	

*Noise* has a smaller effect than *Samples* in both studies. This suggests that the prediction error is more affected by a reduction in sample size than by increased noise. The prediction errors for the two simulation parts are plotted in Fig. 2 and Fig. 3 respectively, and it is clear that all methods have higher prediction errors when the number of samples is low. Also the differences between methods are larger at high noise levels.

As expected, the interaction between *Samples* and *Noise* is quite large, meaning that small data sets with high noise perform even poorer than data set having only low sample size or only high noise.

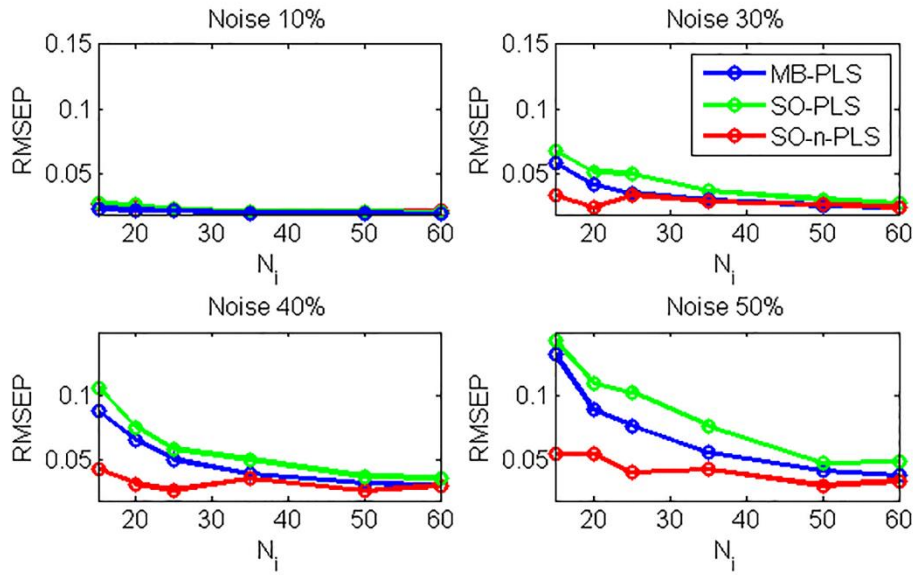


Figure 2: Average RMSEPs for the simulation study Part I. Each subplot shows a different noise level ( $L$ ), while the number of samples ( $N$ ) are given on the abscissa. The curves represent the three methods; SO-N-PLS (red), SO-PLS (green) and MB-PLS (blue).

The averaged RMSEPs for Part I are presented in Fig. 2. The three regression methods show comparable performance for 10% of added noise, and at this noise level the number of samples has little effect on the prediction error. When the noise is higher, SO-N-PLS consistently gives better predictions than the other methods. The difference is largest when the noise level is high and the number of samples is low. These results are in agreement with the initial hypothesis; SO-N-PLS will provide better predictions than unfolded analysis on small sample sizes and on noisy data.

Fig. 3 shows averaged RMSEP values for Part II of the simulation study, where the number of latent variables are chosen to minimize the RMSEP in each model.

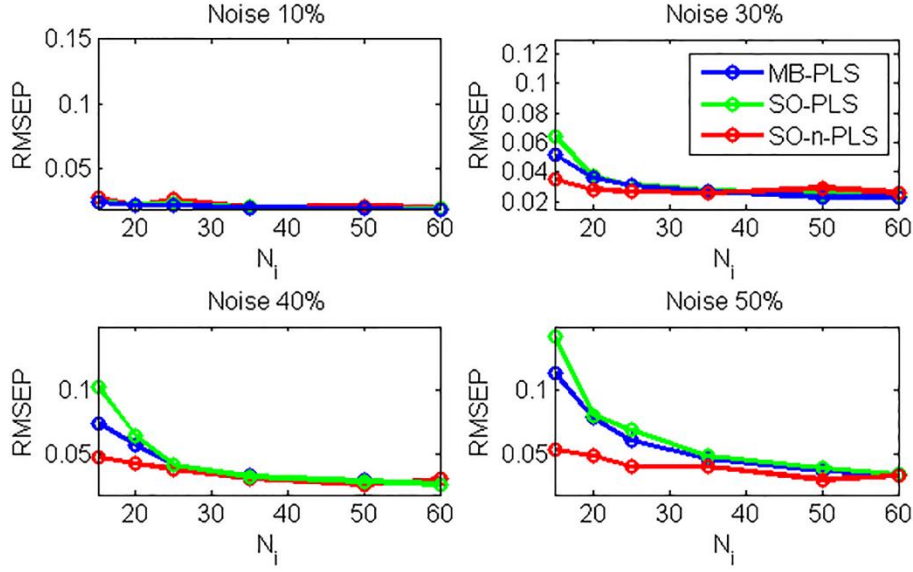


Figure 3: Average RMSEPs for the simulation study Part II. Each subplot shows a different noise level ( $L$ ), while the number of samples ( $N$ ) are given on the abscissa. The curves represent the three methods; SO-N-PLS (red), SO-PLS (green) and MB-PLS (blue).

The results in Fig. 3 can be compared to Fig. 2, and the trends are very similar: SO-N-PLS outperforms the other methods when the noise is high and number of samples is small. Note however that the differences between methods become much smaller when the number of latent variables is selected as part of the modeling. This is in close correspondence with the ANOVA: the differences between methods are smaller in Part II. Here, there is no relevant difference between any of the methods when the number of samples is 25 or more (30-40% noise) and 35 or more (50% noise). The results from SO-PLS and MB-PLS are also comparable, which suggests that SO-PLS and MB-PLS are similar from a prediction point of view when the number of latent components can be adjusted freely.

SO-PLS gives the highest prediction errors in both parts of the simulation study, but in Part II the results were very similar to MB-PLS. This shows that using the “correct” number of components for both blocks is not optimal for SO-PLS. This is likely due to the fact that residuals from the first fit carry information about the noise and the unmodelled structure in  $\mathbf{X}$ . This needs to be corrected when fitting  $\mathbf{Y}$  to the orthogonalized  $\mathbf{Z}$ . As a consequence, SO-PLS could need more components than the “correct” number for the second block.

An additional simulation was run to investigate how SO-PLS and MB-PLS handle noise in  $\mathbf{Y}$ . One hundred data sets were simulated as described above, and 20% noise was added to  $\mathbf{Y}$  each time. SO-PLS and MB-PLS models were fitted both before and after the addition of noise. Results show that SO-PLS gives slightly better predictions than MB-PLS when  $\mathbf{Y}$  is without noise, but these results are reversed when noise is added. This supports the conclusion in the previous paragraph.

Fig. 4 shows the average number of latent variables selected for each level of noise and sample size. SO-N-PLS uses the same number of components as used to generate the data (two components are always chosen) and is therefore not included in Fig. 4. The unfolded methods always select a number of components higher than used to generate the data. MB-PLS (in blue)



generally selects five latent variables for the low noise level, and six when the noise is higher. SO-PLS (in green) generally selects three latent variables for the  $\mathbf{X}$ -block and between four and six latent variables for the  $\mathbf{Z}$ -block (dashed green line). These results suggest that the second initial hypothesis is also valid; SO-N-PLS gives models that are more parsimonious, which is an advantage from the interpretation point of view.

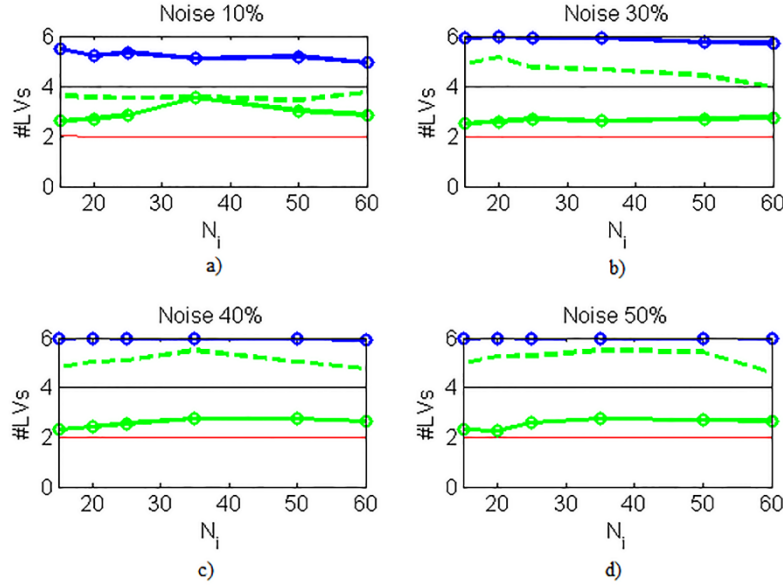


Figure 4: Each subplot shows a different noise level ( $L$ ), while the number of samples ( $N$ ) are given on the abscissa. The curves represent the two methods; SO-PLS (green) and MB-PLS (blue). Red and black continuous lines represent the proper complexity for SO-PLS and MB-PLS, respectively. Dashed lines represent the regression involving the  $\mathbf{Z}$ . SO-N-PLS not shown because two LVs were always selected for both blocks.

In order to further investigate the differences in interpretation, the  $\mathbf{X}$ -weights from SO-PLS and SO-N-PLS on one of the simulated data sets are shown in Fig. 5. The selected data set consists of 60 samples and the noise level is 50% for both  $\mathbf{X}$  and  $\mathbf{Z}$ .

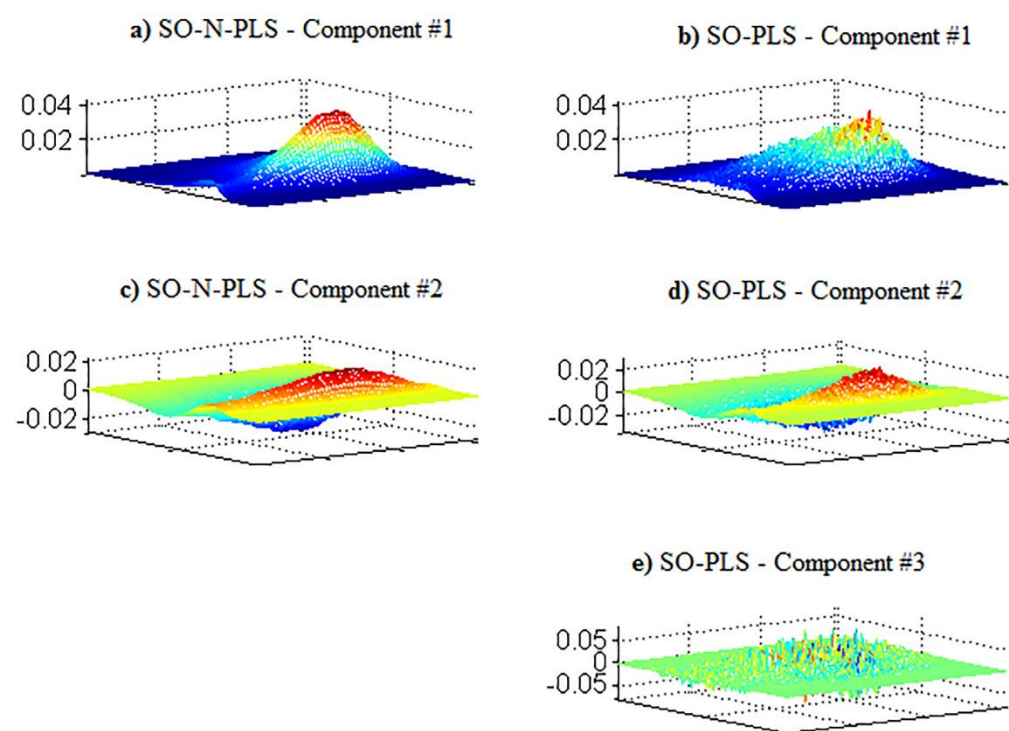


Figure 5:  $\mathbf{X}$ -weights from models on one of the simulated data sets with sixty samples and 50% noise.

Looking at Fig. 5, it is clear that the  $\mathbf{X}$ -weights from SO-N-PLS are smoother than those from the SO-PLS model. Even if the shape of the weights are similar for the two methods, it is evident already in the first component that SO-PLS models more noise than SO-N-PLS. The third component is strongly influenced by noise, which is reasonable since the number of components used to generate the data is two. The same conclusion is reached looking at the  $\mathbf{Z}$ -weights from SO-N-PLS and the  $\mathbf{Z}$ -loadings from SO-PLS, and therefore the plots are not reported here. The same behavior is also observed for lower noise levels.

As explained in paragraph 2.6, the regression coefficients can also be used to graphically interpret models. Regression coefficients from SO-PLS and SO-N-PLS built on a simulated data set (the same data set shown in Fig. 5) are shown in Fig. 6.

These coefficients correspond to  $\mathbf{g}$  and  $\mathbf{y}$  in Eq.1 and Eq.4, respectively. Both sets of coefficients have been refolded to the three-way structure before plotting Fig. 6. The visual appearance confirms that SO-PLS' regression coefficients are more affected by noise than SO-N-PLS'.

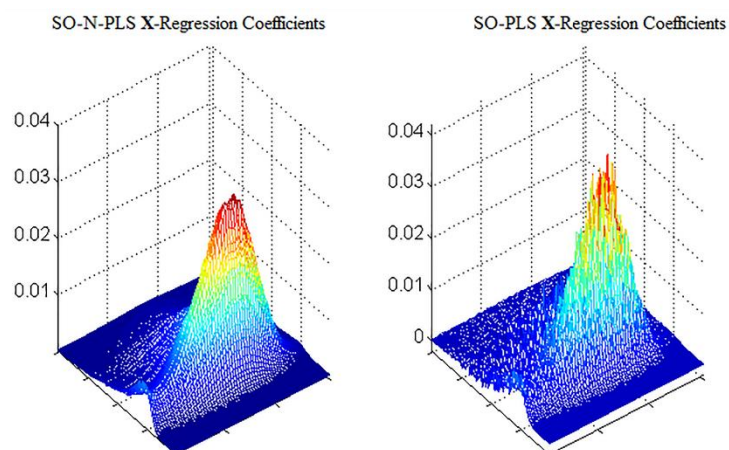


Figure 6: Comparison of X-regression coefficients from SO-N-PLS (left plot) and from SO-PLS (right plot), from a simulated data set with sixty samples and 50% noise.

The results illustrate that SO-N-PLS is better at filtering out noise when an underlying three-way structure is present, while it is included in the model in the unfolded analysis. The plot for MB-PLS is similar and therefore not shown.

#### 4.2 Results on chemical mixture data set

Prediction models for each of the five chemical compounds were fitted by N-PLS using only one block at a time, and by SO-N-PLS, MB-PLS and SO-PLS on both blocks. Results are reported in Table 2.

Table 2: Chemical mixtures data set: RMSECVs and Explained variances for the prediction of the concentrations of compounds in the mixture.

Compound	N-PLS (Only X-block)			N-PLS(Only Z-block)			SO-N-PLS			MB-PLS			SO-PLS		
	LVs	RMSCV	Expl. Var. (%)	LVs	RMSCV	Expl. Var. (%)	LVs	RMSCV	Expl. Var. (%)	LVs	RMSCV	Expl. Var. (%)	LVs	RMSCV	Expl. Var. (%)
Valine-Tyrosine-Valine	3	0.59	97	2	0.74	96	2,2	0.19	99	3	0.19	99	2,4	0.14	99
Tryptophan-Glycine	2	0.20	99	2	0.86	96	2,2	0.12	100	3	0.15	99	3,5	0.11	100
Phenylalanine	2	0.30	98	2	0.91	95	1,2	0.26	99	3	0.26	99	3,4	0.21	98
Maltoheptaose	1	2.00	84	2	0.15	99	1,2	0.14	99	2	0.14	99	3,2	0.12	100
Propanol	1	2.22	84	2	0.7	96	2,2	0.51	97	5	0.20	99	1,5	0.09	100

Explained variances for all compounds are generally high for at least one of the one-block models, but were sometimes slightly improved by using both blocks. These differences are more evident when looking at the RMSECVs. In all cases, the SO-PLS model gave the best prediction results. SO-N-PLS results were comparable except from one case (propanol) where the difference is large. The MB-PLS predictions were in all cases less precise than SO-PLS.

These results are not in accordance with the hypothesis, since the unfolded methods perform better than SO-N-PLS. A thorough examination of the model revealed that the reason possibly stems from the handling of non-linearity in the data. Fig. 7 shows the concentration of propanol as a function of the first two  $\mathbf{Z}_{orth}$ -components from SO-PLS and SO-N-PLS. The second component from SO-PLS is the only one that has a clear, linear relationship with propanol. This suggests that SO-PLS due to its flexibility may be better at finding relevant linear combinations of the data than the SO-N-PLS which always obeys a tree-way data structure. The high number of components SO-PLS selected from the  $\mathbf{Z}$ -block also confirms this hypothesis.

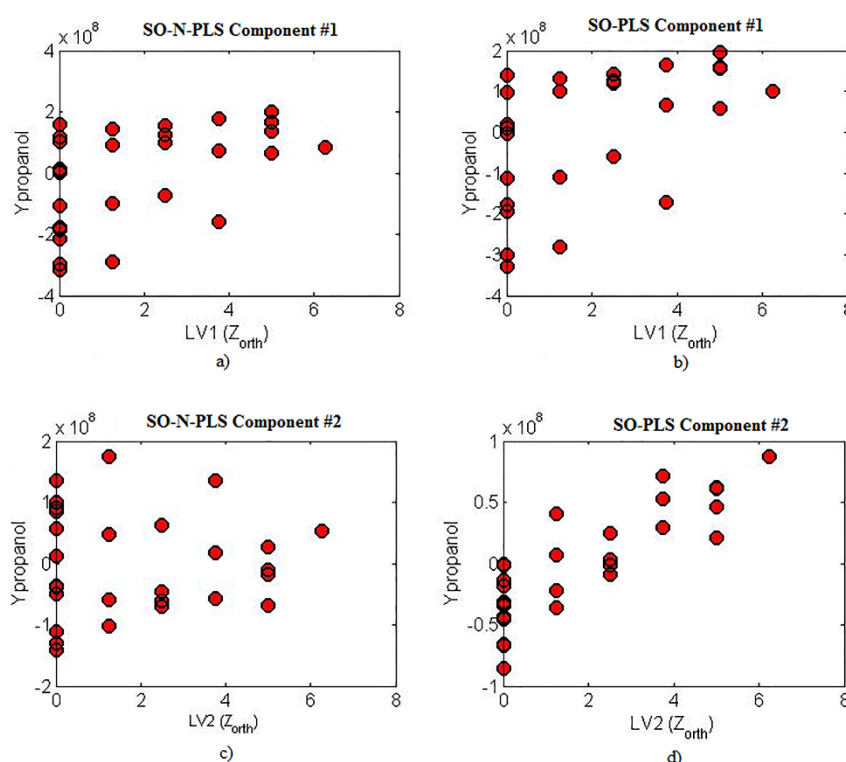


Figure 7:  $y_{Pro}$  vs  $\mathbf{T}_{Z_{orth}}$  from the SO-N-PLS and SO-PLS models: a) and b)  $y_{Pro}$  is reported as a function of the first  $\mathbf{T}_{Z_{orth}}$  from SO-N-PLS and the first  $\mathbf{T}_{Z_{orth}}$  from SO-PLS, respectively. c) and d)  $y_{Pro}$  plotted against the second  $\mathbf{T}_{Z_{orth}}$  from SO-N-PLS and from SO-PLS, respectively.

The methods differ slightly in the number of selected latent variables. In general, MB-PLS selects less components than SO-PLS, which selects the highest number (among the three methods). SO-N-PLS selects less components than MB-PLS only in one case (two plus two components selected for SO-N-PLS and five for MB-PLS). In another case, the two methods select the same number of latent variables (one plus two and three); in all the other models SO-N-PLS selects one component more than MB-PLS. This is different from the simulation study, where SO-N-PLS gives the most parsimonious models. This could also be due to the presence of non-linearity, SO-(N)-PLS needs more components to handle it. Alternatively, it could stem from the fact that in the simulations, the response was affected by independent components from both  $\mathbf{X}$  and  $\mathbf{Z}$ , while the relevant information might be overlapping in this data set.

The aim of this study is not to give a detailed chemical interpretation of the system, but rather to highlight differences between the graphical interpretations of the methods. As an example, weights from the SO-N-PLS, SO-PLS and the MB-PLS models (related to the **X**-block) for the prediction of *Valine-Tyrosine-Valine* are reported in Fig. 8. For SO-N-PLS, the outer product of the second and third mode weights is plotted. For MB-PLS and SO-PLS, weights are refolded before being plotted. As mentioned in the initial hypothesis, models built with a small number of components are easier to interpret. Note that, even if MB-PLS has the lowest number of latent variables (three versus two plus two) we need to interpret three plus three weights for MB-PLS (three latent variables correspond to three components per each block) versus two plus two for SO-N-PLS. Consequently, SO-N-PLS model has the least number of weights to plot and interpret.

The **X**-weights from the SO-N-PLS model show that the two components represent two different compounds, one that has emission around 300 nm and excitation around 275; and the other that has emission around 280 nm and excitation around 260 nm. Looking at the fluorescence spectra of the pure compounds, these correspond to *Valine-Tyrosine-Valine* and *Phenylalanine*, respectively. The same interpretation is not so straightforward from the SO-PLS loadings weights. The first component (Fig. 8b) is similar to SO-N-PLS', but the negative peak has a wider shape. This makes the identification of the excitation peak more difficult. In the second component (Fig. 8e) it would be possible to identify the excitation peak, but the emission one is too wide to make a clear interpretation. Peaks identification looks even more difficult for MB-PLS (Fig. 8c, 8f and 8g). Due to the wide shape of the peaks, component one from MB-PLS is difficult to interpret. Components two and three are similar to components one and two (respectively) from the SO-(N)-PLS models; but even in this case the width of the peaks would make the chemical interpretation weak. Note that MB-PLS is in general more complicated to interpret, since each component presents contributions from both blocks.

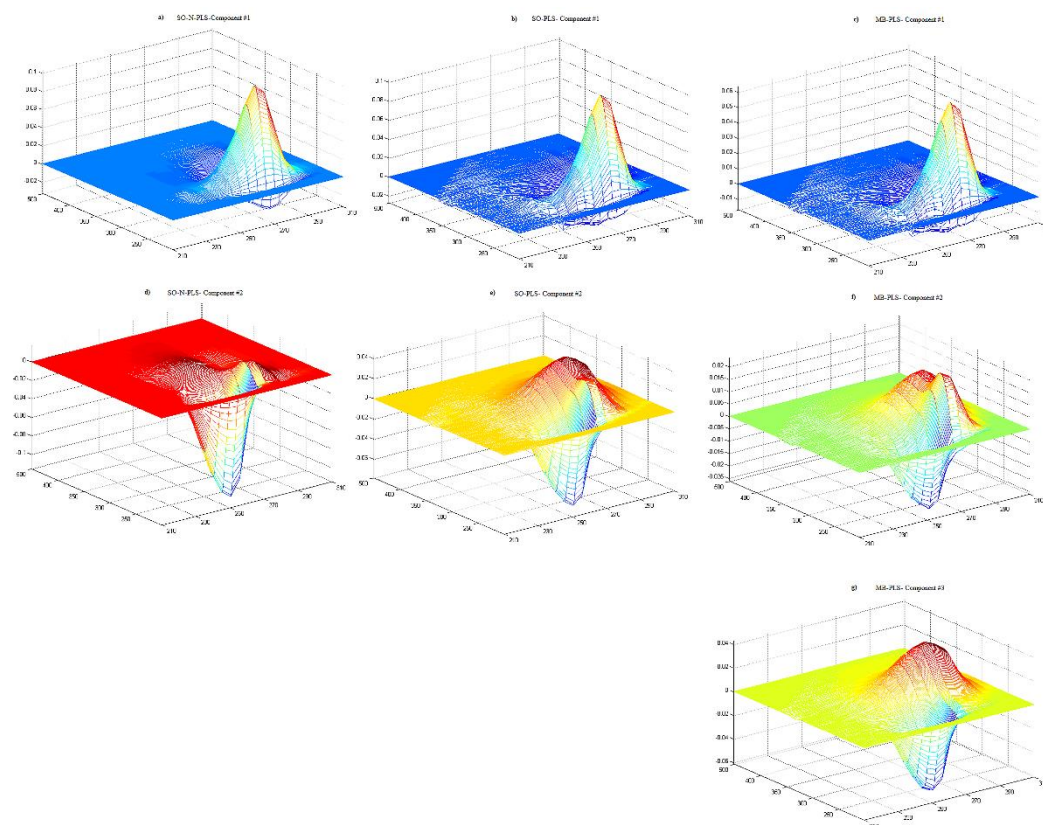


Figure 8: Chemical mixture data set. Weights (for SO-N-PLS, SO-PLS and MB-PLS) plots (related to the **X**-block) for prediction of the *Valine-Tyrosine-Valine* compound.

In order to check the starting hypothesis, a further investigation has been conducted on the chemical data set. Some random noise (simulated as it is described in 3.1 for the simulation study, and correspondent to the 50% of the signal of each block) was added to each predictor and new SO-N-PLS, SO-PLS and MB-PLS models were built. This was replicated ten times, and averaged RMSECVs and selected number of latent variables are reported in Table 3.

Table 3: Averaged (over the 10 replicates) RMSECVs and number of components from SO-N-PLS, MB-PLS and SO-PLS models after the addition of 50% random noise to  $\mathbf{X}$  and  $\mathbf{Z}$ .

Compound	SO-N-PLS			MB-PLS		SO-PLS		
	RMSCV	# components		RMSCV	# components	RMSCV	# components	
		X	Z				Z	Z
Valine-Tyrosine-Valine	0.31	1.3	2.0	0.32	4.0	0.50	1.5	5.1
Tryptophan-Glycine	0.13	1.2	2.0	0.49	3.8	0.22	3.3	5.0
Phenylalanine	0.24	1.4	2.0	0.38	4.1	0.25	1.9	2.9
Maltoheptaose	0.14	1.4	2.0	0.15	3.3	0.17	1.9	5.0
Propanol	0.72	1.5	2.0	0.55	4.1	0.75	1.0	5.4

The averaged RMSECVs from the new models agree with the results from the simulation study, supporting the first hypothesis. Except for the prediction of the propanol, SO-N-PLS is giving the lowest RMSECVs. For chemical reasons, the use of fluorescence spectra ( $\mathbf{X}$ -block) to predict propanol cannot be completely reliable from the analytic point of view and its prediction cannot be consider an indicator of the model performances. Consequently, SO-N-PLS is confirmed as the best predictor method (among these three) for noisy data.

Concerning the number of latent variables, SO-N-PLS is once again the most parsimonious method in selecting latent variables. These results are also in agreement with the simulation study, supporting the second hypothesis.

#### 4.3 Results on the Lambrusco data set

##### *Classification results*

Classifications of Lambrusco wines were first performed by single block methods; N-PLS-LDA on the three-way  $\mathbf{X}$  (GC-MS) and PLS-LDA on the two-way  $\mathbf{Z}$  (NMR). Then, these models were compared to the multi-block methods MB-PLS-LDA, SO-PLS-LDA and SO-N-PLS-LDA. Results for all models are reported in Table 4. It is clear that the  $\mathbf{X}$ -block has the highest discriminating power, giving a total classification error of 24% versus 59% for the  $\mathbf{Z}$ -block. By combining  $\mathbf{X}$  and  $\mathbf{Z}$ , the error is unchanged for SO-N-PLS-LDA and SO-PLS-LDA and one sample more is misclassified by MB-PLS. In other words, the multi-block models gave almost identical results to the model using only  $\mathbf{X}$ . The numbers of latent variables are the same for all multi-block models: six for MB-PLS and two plus four for SO-(N)-PLS.

Table 4: Lambrusco Data set: Classification errors by single-block and multi-block methods.

Method	LVs:	Miscl. Grasparossa	Misclassified Salamino	Misclassified Sorbara	Tot.Error (%)
N-PLS (Only X)	3	7	4	3	24
PLS (Only Z)	3	16	13	5	59
SO-N-PLS-LDA	2,4	7	4	3	24
MB-PLS-LDA	6	6	7	2	26
SO-PLS-LDA	2,4	6	5	3	24

One way to interpret the SO-N-PLS-LDA models is to look at the cross-validated predictions in the space of the canonical variates, as shown in Fig. 9. In order to do that, the cross-validated  $\mathbf{Y}$ -values are used to calculate the covariance matrix necessary to extract the canonical variates. More details can be found in [15].

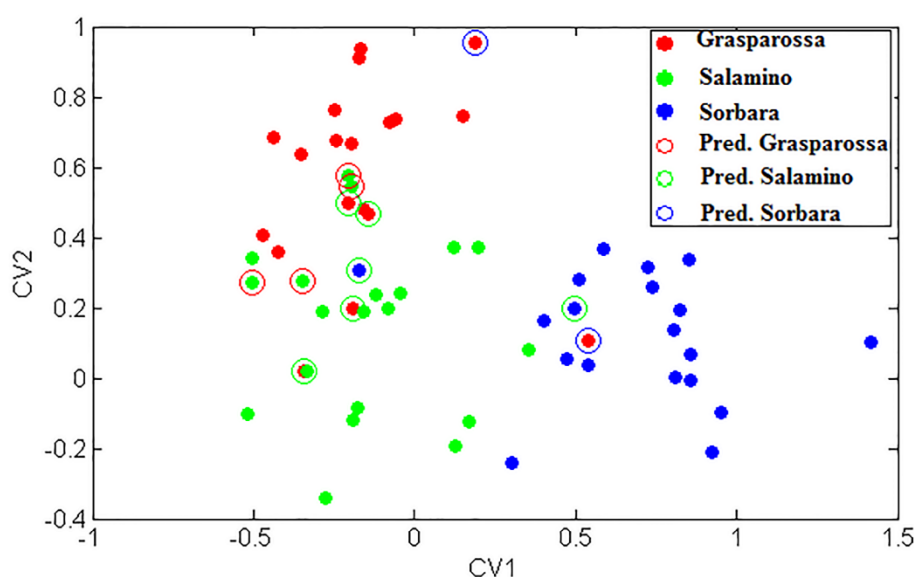


Figure 9: Classification of Lambrusco wines. Predictions in the space on canonical variates using both  $\mathbf{X}$  (GC-MS) and  $\mathbf{Z}$  (NMR) blocks. Circled samples are the misclassified ones.

There is a strong overlap between the Grasparossa and Salamino classes. The reason for this is that both wines are made from mixtures. According to law, *Lambrusco Salamino di Santa Croce PDO* contains 85% of Salamino grape and the rest 15% is of grapes harvest in Modena's area (so they could be Grasparossa or Sorbara). The same applies to "*Lambrusco Grasparossa di Castelvetro PDO*", while "*Lambrusco di Sorbara PDO*" contains 60% of Sorbara grape plus 40% Salamino grape.

In order to focus on the differences between Salamino and Sorbara, and to check the possibility of distinguishing between the two, a new classification models were fitted only to the thirty-nine Salamino and Sorbara samples. Cross-validated predictions in the space of the canonical variates (from the SO-N-PLS-LDA model) are visualized in Fig. 10. Some misclassification cannot be



avoided due to the nature of wines: two Salamino and three Sorbara samples are misclassified in this model (Fig. 10, red bars), and the classification error is 10% and 16% for Salamino and Sorbara respectively. N-PLS-LDA on  $\underline{X}$ , MB-PLS-LDA and SO-PLS-LDA misclassify the same samples, indicating that these are intrinsically hard to distinguish. PLS-LDA on  $\underline{Z}$  misclassifies even more samples (six and three misclassified for Salamino and Sorbara, respectively).

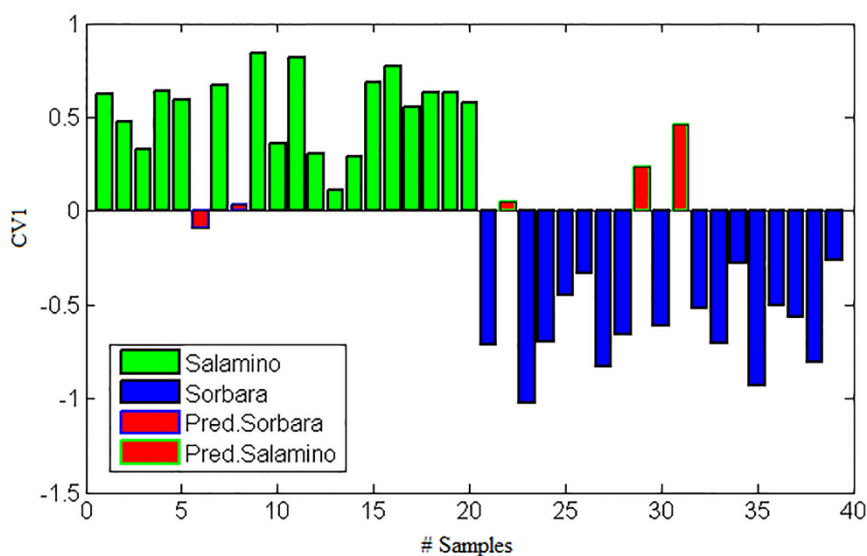


Figure 10: Classification of Lambrusco wines: Cross-validated predictions in the CVA space using both blocks restricted to only two classes (Salamino and Sorbara). Red bars are the misclassified samples.

#### 4.4 Discussion of the hypotheses mentioned in the introduction

The first hypothesis was that SO-N-PLS is expected to give better predictions for small sample sizes and noisy data. The simulation study confirms this, since SO-N-PLS performs better than the unfolded methods except when the noise is low (10%). For the low noise level, the three methods are comparable regardless of sample size. SO-N-PLS also filters the noise better than the other methods, which is clearly seen in Fig. 5 and 6. SO-N-PLS outperforms the other regression methods in particular when the number of components is set equal to the true number (Part I of the simulation). In all cases, the difference between SO-N-PLS and SO-PLS is higher than the difference between SO-N-PLS and MB-PLS. In the more realistic scenario where the number of components is determined by cross-validation (Part II of the simulation), the difference between MB-PLS and SO-PLS became negligible.

The superiority of SO-N-PLS is not visible in the real data sets, which is probably due to non-linearities and less clear three-way structure in data. In the chemical mixtures data set, the SO-PLS was the best in prediction. Nevertheless, in the further study made on this data set, the behaviors shown in the simulation study are visible again. In fact, after the addition of random noise to  $\underline{X}$  and  $\underline{Z}$ , SO-N-PLS gives the best predictions. Concerning predictions in the Lambrusco data set, the methods were indistinguishable. In the chemical mixtures data set, multi-block methods improved predictions slightly as measured by RMSECV, while in the Lambrusco data set they did not. It is important to mention, however, that for practical use of the methods these



results should be validated more carefully using a new test set, the reason being that the both selection of components and the actual prediction results are based on the same cross-validation.

The second hypothesis was that SO-N-PLS leads to simpler models that could be more easily interpreted. This is not completely confirmed, but some clear indications are given in the simulation study. In the simulations, SO-N-PLS always selects the actual underlying complexity, while MB-PLS and SO-PLS generally need more latent variables (in particular for the **Z**-block). This may lead to less stable predictions and more model parameters (weights) to interpret.

Looking at the real data sets, this overestimation of latent variables by MB-PLS and SO-PLS is less evident. For the mixture data set, MB-PLS needs less latent variables than SO-N-PLS. Nevertheless, MB-PLS leads to a more complicated interpretation since all the parameters that have to be investigated are doubled (each component gives loadings for both blocks). SO-PLS still needs more latent variables than SO-N-PLS. In the further study on this data set with addition of noise to **X** and **Z**, SO-N-PLS confirms its parsimony in the latent variable selection.

Considering the graphical interpretation of the models, SO-N-PLS' weights can be represented directly or mode-wise. Anyhow, comparable plots of weights and regression coefficients can be made based on all three methods. In these, we have shown that SO-N-PLS is better at filtering out noise and thereby gives more clear/interpretable plots.

## 5. Conclusions

The novel method SO-N-PLS can be used to fit multi-block models when predictor blocks are multi-way arrays, without unfolding the arrays. The method can be applied for both prediction and classification. It shows some benefits when compared to methods based on unfolded data (SO-PLS and MB-PLS), given that the three-way data satisfies a low-dimensional three-way structure.

Simulation studies showed that SO-N-PLS performs better than the unfolded methods when the sample size is small and the data is noisy. This is due to the fact that it filters out the noise better than MB-PLS and SO-PLS. For the two real data examples, the superiority of SO-N-PLS method is not so evident, but it performed well also for these cases.

SO-N-PLS has many of the same properties as SO-PLS: it is invariant to block scaling and it allows for different numbers of components for each block. It also has some benefits related to interpretation, since the contribution from each block can be interpreted individually. Additionally, SO-N-PLS gives rise also to a number of graphical interpretation tools. The advantage of these is that they take into account the original three-way structure of the data.

## 6. Acknowledgements

We would like to thank the Norwegian Levy on Agricultural Products (FFL) and the Research Council of Norway for financial support (Project number: 225096).

## 7. References

- [1] R. Coppi, S. Bolasco, Multiway Data Analysis. Amsterdam: North-Holland, 1989.
- [2] R. Bro, Multi-way analysis in the food industry: models, algorithms, and applications. Amsterdam: Universiteit van Amsterdam, 1998 .
- [3] P.M. Kroonenberg, Applied Multiway Data Analysis. Wiley Series in Probability and Statistics 702. John Wiley & Sons, New York, 2008.
- [4] R.A. Harshman, Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis, UCLA Working Papers in phonetics, 16 (1970) 1.
- [5] J. D. Carroll, J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young decomposition, Psychometr. 35 (1970) 283.
- [6] R. Bro, Multiway calibration. Multilinear PLS, J. Chemometr. 10 (1996) 47-61.
- [7] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometr. 31 (1966) 279–311.
- [8] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometr. 10 (1996) 463–482.
- [9] J.A.Westerius, T. Kourti, J.F. MacGregor, Analysis of hierarchical PCA and PLS models, J. Chemometr. 12 (1998) 301-321.
- [10] K. Jørgensen, V. Segtnan, K. Thyholt, T. Næs, A comparison of methods for analysing regression models with both spectral and designed variables, J. Chemometr. 18 (2004) 451–464.
- [11] T. Næs, O. Tomic, B. H. Mevik, H. Martens, Path modelling by sequential PLS regression, J. Chemometr. 25 (2011) 28–40.
- [12] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: Separating common and unique information in several data blocks, Food Qual. Pref. 24 (2012) 8–16.
- [13] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemometr. 25 (2011) 441–455.
- [14] E. Acar, T. G. Kolda, D. M. Dunlavy, All-at-once Optimization for Coupled Matrix and Tensor Factorizations, MLG'11: Proceedings of Mining and Learning with Graphs (2011).
- [15] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, Chemometr. Intell. Lab. Syst. 141 (2015) 58–67.
- [16] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188.
- [17] T. Næs, U. Indahl, A unified description of classical classification methods for multicollinear data, J. Chemometr. 12 (1998) 205–220.
- [18] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, New York, 1989.
- [19] S. De Jong, Short communication regression coefficients in multilinear PLS, J. Chemometr. 12 (1998) 77-81.
- [20] R. Bro, PARAFAC: Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (1997) 149-171.
- [21] E. Acar, E. E. Papalexakis, G. Gurdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, R. Bro, Structure-Revealing Data Fusion, BMC Bioinformatics, 15 (2014).
- [22] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, Mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, Chemometr. Intell. Lab. Syst. 137 (2014) 181–189.
- [23] U. Indahl, A twist to partial least squares regression, J. Chemometr. 19 (2005) 32–44.

## **Paper IV**

**Title:** Multiblock regression: combining data sets with different structures and dimensionalities

**Authors:** A. Biancolillo, T. Næs, R. Bro, I. Måge

*Manuscript. It will be submitted to Chemometrics and Intelligent Laboratory Systems tentative date March 2017*

# Multiblock regression: combining data sets with different structures and dimensionalities

Alessandra Biancolillo <sup>a,b,\*</sup>, Tormod Næs<sup>a,b</sup>, Rasmus Bro<sup>b</sup>, Ingrid Måge<sup>a</sup>

<sup>a</sup> *Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway*

<sup>b</sup> *Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

\*Corresponding author:

Tel: +47 64 97 01 15

e-mail: [alessandrabiancolillo@hotmail.it](mailto:alessandrabiancolillo@hotmail.it); [alessandra.biancolillo@nofima.no](mailto:alessandra.biancolillo@nofima.no)

**Keywords:** Multi-block regression; SO-PLS; MB-PLS; PO-PLS; components

## Abstract

Data-fusion is very useful for the extraction of information from multi-block data, but the choice of the most appropriate multi-block method is not straightforward. Which method would be the most suitable depends on the data and on the purpose of the analysis. In this work, the behavior of three multi-block methods handling data sets which present different underlying dimensionality and, at the same time, a different number of variables, is investigated. The discussion will be mainly focused on interpretation rather than on predictions. The three multi-block methods are MB-PLS, SO-PLS and PO-PLS; they will be applied on simulated data sets. In one part of the simulation study, multi-block methods are used to combine process-like and spectra-like data. In the second part, the same three methods handle a categorical and a spectra-like block. Special attention will be given to the interpretation of the loadings; in particular, a method to investigate their interpretability is proposed. Additionally, the selection of the number of components in the different data structures will be discussed. In particular, the agreement between the “expected” number of components (namely, the actual inner dimensionality of the data) and the optimal complexity required by the models will be discussed.

## 1. Introduction

Many methods for handling multi-block regression problems have been developed, mainly driven by advances in sensor technologies [1,2]. The choice of method depends on both the objective of the data analysis and the nature of the data sets. Previous work has shown that different methods usually give models with similar predictive performance, but the interpretation of model parameters differs and is not straightforward [2-7]. Therefore, there is a strong need to investigate the interpretation of these models further.

The interpretation might depend on method characteristics such as the selection of components and how the model parameters are calculated, as well as on the properties of the data blocks. In that regard, it is of especial interest to investigate how data blocks with different sizes and covariance structures are handled.

In this paper, three multi-block regression methods based on Partial Least Squares (PLS) regression are compared and discussed from the interpretation point of view. The methods are Multiblock-PLS (MB-PLS) [8-11], Sequential Orthogonalized PLS (SO-PLS) [12-13] and Parallel Orthogonalized PLS (PO-PLS) [14-15]. MB-PLS is a well-established method which is easy to implement for any number of blocks, since it essentially is a regular PLS regression on the concatenated blocks. SO-PLS and PO-PLS were developed to give better interpretation for specific data analytical objectives: SO-PLS fits the blocks sequentially, which allows interpretation of the incremental information in each block. PO-PLS, on the other hand, identifies common and distinct components in the data blocks and is therefore more suited for interpreting the correspondence between blocks. Both SO-PLS and PO-PLS allow for different numbers of components in each block, which can be seen as an advantage from an interpretation point of view. The hypothesis of this paper is therefore that SO-PLS and PO-PLS provide a better insight into the underlying phenomena in the data set, especially when the blocks have different sizes and covariance structures.

## 2. Materials and methods

All the methods used in this paper can handle any number of predictor blocks, but we will focus on the two-block case represented by the equation:

$$Y = X\beta + Z\gamma + E \quad (1)$$

Where  $Y_{(N \times I)}$  is the response,  $X_{(N \times J)}$  and  $Z_{(N \times K)}$  are the predictor blocks and  $E_{(N \times I)}$  is the residual matrix. Both predictors and responses are assumed to be mean-centered (when not differently specified).

In the following we present briefly the three methods considered in the paper. In Section 2.5 we discuss more deeply how to interpret the models.

### 2.1 MB-PLS

In Multi-Block-PLS [8-11], the concatenated matrix  $X_{conc} (X_{(N \times (J+K))}^{conc} = [X \ Z])$  is used to predict the response  $Y$  by PLS regression. In order to avoid that blocks with several variables drive the regression, predictor blocks need to be block-scaled before the concatenation. In this work, block-scaling is carried out dividing each block by its Froebenius norm.

### 2.2 SO-PLS

As the name suggests, Sequential and Orthogonalized-PLS [12-13] is a multiblock regression method where the information from the different predictors are extracted sequentially to achieve the final model. The algorithm is divided in four steps:

- 1)  $Y$  is fitted to  $X$  by PLS, obtaining scores  $T_X$
- 2)  $Z$  is orthogonalized with respect to the scores of PLS in 1 ( $T_X$ )
- 3) Residuals from 1 are fitted to  $Z_{orth}$  by PLS, obtaining  $T_{Zorth}$
- 4) The final model is calculated as in Eq. (1)

This method has a number of properties that make it particularly suitable for interpretation purposes. The number of components can be defined for each block involved in the regression. Moreover, the contribution of each block can be evaluated individually. The investigation of each PLS-regression involved in the model allow understanding which info come from which block. More details can be found in [12-13].

### 2.3 PO-PLS

PO-PLS [14-15] distinguishes between *common* and *distinctive* components in the predictor blocks. The PO-PLS algorithm applied for this work is not the original one [14] but its variation presented in [15]. It can be summed up in the following steps:

- 1)  $Y$  is predicted from  $X$  and  $Z$  by two individual PLS models. Scores from this model are called  $T_X$  and  $T_Z$ .
- 2) Common components are identified by canonical correlation analysis [16-17] of  $T_X$  and  $T_Z$ , and the *number* of common components is decided by evaluating the canonical correlations and the explained variances in each block. The common scores  $T_C$  ( $C$  for common) are then defined as the average canonical scores from each block, and  $Y$  is fitted to  $T_C$  by ordinary least squares regression.
- 3) Scores  $T_X$  from 1) are orthogonalized with respect to  $T_C$ , giving  $T_{Xorth}$ , and  $Y$  is then predicted from  $T_{Xorth}$  by PLS regression giving  $T_{DXorth}$  ( $D$  for distinct). These scores represent the distinct information in  $X$ .
- 4) Scores  $T_Z$  from 1) are orthogonalized with respect to  $T_C$  and  $T_{DXorth}$ , giving  $T_{Zorth}$ , and  $Y$  is then predicted from  $T_{Zorth}$  by PLS regression giving  $T_{DZorth}$ . These scores represent the distinct information in  $Z$ .
- 5) The final predictive model is obtained in the same way as for SO-PLS by regressing  $Y$  on  $T_{PO} (= [T_C \ T_{DXorth} \ T_{DZorth}])$ . If there are no common components (in step 2) the model is identical to SO-PLS.

## 2.4 Choosing the optimal number of components in MB-PLS, SO-PLS and PO-PLS

The number of components needs to be selected for each method. For MB-PLS, there is only *one* number of components to choose, which can be decided by e.g. evaluating the cross-validated root mean squared errors (RMSECV). In SO-PLS one needs to select the number of components for each block, i.e. two numbers for the model in Eq. 1. In PO-PLS one needs to select the number of common components as well as the number of distinctive components in each block, which amounts to *three* numbers for the model in Eq. 1.

In SO-PLS and PO-PLS, selection of the optimal complexity can be done in two ways: the *sequential* or the *global approach* [12]. In the sequential approach, one starts by selecting the optimal number of components for the first regression. Once the optimal complexity is fixed for the first regression, the number of components to be used in the second one is chosen. In the global approach, all possible combinations of components (below a fixed maximal amount) are used to build models. In both approaches, the optimal complexity is defined looking at the root mean square error in cross-validation (RMSECV). In order to ease the selection in the global approach, the so-called *Måge plot* [12] can be investigated. In this, RMSECVs from the different models are shown as function of the total number of components. In real applications, it is advisable to manually inspect Måge plots and select the optimal number of components as a tradeoff between model size and predictive ability. The optimal number is often not the absolute minimum. In simulation studies is not possible to look at plots for all the repetitions, and we therefore define the optimal number of components as the smallest number of components giving an RMSECV not significantly different from the absolute minimum (Each RMSECV is tested by a  $\chi^2$  test with 5% significance level [18]). If none of the RMSECVs in the error distribution is comparable to the absolute minimum, then the number of components corresponding to the lowest RMSECV is chosen.

Since MB-PLS involves only one regression on the concatenated predictors, only one number of components has to be selected. Consequently, it could be that the optimal complexity for  $X_{conc}$  is not reflecting the optimal complexity for each individual block, leading to an

overestimation/underestimation of its components. Obviously, this affects the interpretation of the contribution of each block to the model. Instead, in PO- and SO-PLS, the number of components is defined independently per block. Therefore, it is possible to extract from each block exactly the number of components that is considered to be optimal for it.

## 2.5 Interpretation of multiblock models

MB-, SO- and PO-PLS have different model parameters that can be investigated for interpretation purposes. For all these methods, parameters such as scores, loadings and regression coefficients can be interpreted as in regular PLS.

Interpretation of loadings in SO-PLS and PO-PLS is slightly different than in PLS. In particular regarding the interpretation of  $\mathbf{Z}$ -loadings in SO-PLS and  $\mathbf{X}$ - and  $\mathbf{Z}$ -loadings in PO-PLS. Due to the orthogonalization step, these loadings require a slight modification before being interpreted. This has been shown in a previous work [19] for  $\mathbf{Z}$ -loadings in SO-PLS; the same reasoning applies also to  $\mathbf{X}$ - and  $\mathbf{Z}$ -loadings in PO-PLS. A matrix and its orthogonalized counterpart do not necessarily belong to the same row space; this means that  $\mathbf{X}_{orth}$  and  $\mathbf{X}$  (in PO-PLS) and  $\mathbf{Z}_{orth}$  and  $\mathbf{Z}$  (in both SO- and PO-PLS) may be in different row spaces. When interpreting these models, the original  $\mathbf{Z}$  (and  $\mathbf{X}$ ) can be projected down onto the orthogonalized scores  $\mathbf{T}_{Zorth}$  (or  $\mathbf{T}_{Xorth}$ ), obtaining “recalculated” loadings  $\mathbf{P}_Z$  for interpretation [19]:

$$\mathbf{P}_Z = (\mathbf{T}_{Zorth}^T \mathbf{T}_{Zorth})^{-1} \mathbf{T}_{Zorth}^T \mathbf{Z} \quad (2)$$

These loadings will then be in the same row space as the original data.

In the present paper, high relevance is given to the interpretation of loadings from MB-, SO- and PO-PLS models based on simulated data sets. One of the advantages of simulation studies lies in the possibility of comparing “true” model parameters with the estimated ones. It is, however, not straightforward to define an objective measure of interpretability. We will therefore propose a way of quantifying how well the estimated loadings represent the space spanned by the true (simulated) loadings ( $\mathbf{\Pi}$ ). This approach can be used to have an indication about each method’s “interpretability” in the simulation study. We are going to refer to this as *explained variance criterion*. Note that it is conceived to fit well with simulations (in order to investigate the methods’ behavior) but it is not applicable to models based on real data.

Once multi-block regression models have been built (on simulated data sets), the procedure consists of only three steps:

1. **Loadings extracted from regression models are collected in a matrix** (with  $J$  or  $K$  rows and # of components columns).
2. **The loadings matrix obtained in step 1 is regressed on the true loadings  $\mathbf{\Pi}$**  by ordinary least squares.
3. **Explained variances can be investigated as an index of the goodness of how well the models span the right variable subspaces.** Hence, to some extent it gives an indication on how well the interpretability of the estimated loadings can be trusted.

In order to avoid confusion, the variance explained by the OLS-regression in step 2 will be called *variance span*.

For the geometrical issue exposed at the beginning of the paragraph, applying the explained variance criterion to  $\mathbf{Z}$ -loadings from SO-PLS, two additional steps (step 1b and 1c) can be added between step 1 and step 2 in the previous list:

1b. True  $\mathbf{Z}$ -scores  $\mathbf{\Theta}_Z$  are orthogonalized with respect to the true (simulated)  $\mathbf{X}$ -scores ( $\mathbf{\Theta}_X$ ):

$$\mathbf{\Theta}_{Z\text{Orth}} = \mathbf{\Theta}_Z - \mathbf{\Theta}_X(\mathbf{\Theta}_X^T \mathbf{\Theta}_X)^{-1} \mathbf{\Theta}_X^T \mathbf{\Theta}_Z \quad (3)$$

1c. True  $\mathbf{Z}$ -loadings  $\mathbf{\Pi}_Z$  are projected onto the true (simulated) orthogonalized  $\mathbf{Z}$ -scores ( $\mathbf{\Theta}_{Z\text{Orth}}$ ).

$$\mathbf{\Pi}_Z = (\mathbf{\Theta}_{Z\text{Orth}}^T \mathbf{\Theta}_{Z\text{Orth}})^{-1} \mathbf{\Theta}_{Z\text{Orth}}^T \mathbf{Z} \quad (4)$$

Applying the explained variance criterion to the  $\mathbf{X}$ - and  $\mathbf{Z}$ -loadings from PO-PLS,  $\mathbf{\Theta}_{X\text{Orth}}$  are orthogonalized with respect to the true (simulated) common scores ( $\mathbf{\Theta}_C$ ) in the same way it is done in Eq. 3 ( $\mathbf{\Theta}_{X\text{Orth}} = \mathbf{\Theta}_X - \mathbf{\Theta}_C(\mathbf{\Theta}_C^T \mathbf{\Theta}_C)^{-1} \mathbf{\Theta}_C^T \mathbf{\Theta}_X$ ) and then the true loadings  $\mathbf{\Pi}_X$  are projected onto the true orthogonalized scores  $\mathbf{\Theta}_{X\text{Orth}}$ .

In the previous lines it is suggested that, for simulations, the explained variance criterion should be operated by investigating the loading matrices after projection of the true loadings to the orthogonalized scores. This is different from what is suggested for the interpretation of loadings (Eq.2) [19]. The reason of this difference is that, in this case, it is possible to interpret the extracted loadings as they are (without being projected back in the original space). This gives the opportunity to interpret the individual contributions of the  $\mathbf{X}$ -block (for PO-PLS) and the  $\mathbf{Z}$ -blocks (for both PO- and SO-PLS). In fact, the redundant information has been accounted in the previous PLS-regression and it has been removed by orthogonalization. This is definitely in a good agreement with the philosophy behind the two methodologies.

However, it is also possible to do not deviate too much from the standard procedure proposed in [19] applying the explained variance criterion comparing the “true” simulated  $\mathbf{X}$ -loadings (for PO-PLS) and  $\mathbf{Z}$ -loadings (for both PO-PLS and SO-PLS) with the correspondent extracted loadings projected back in the original space (as indicated in Eq 2). For the back projection of the extracted  $\mathbf{X}$ -loadings from PO-PLS, equation is:  $\mathbf{P}_X = (\mathbf{T}_{X\text{Orth}}^T \mathbf{T}_{X\text{Orth}})^{-1} \mathbf{T}_{X\text{Orth}}^T \mathbf{X}$ .

Indeed, a simulation study carried out for the sake of confirming this intuition, demonstrated that the results of the two approaches are in good agreement.

Three different scenarios can arise:

1. Models can be built with the theoretically “true” number of components (according to the simulation parameters) and they can present shapes of loadings similar to the originals. These models will have:
  - *Variance span* close to 100% for all estimated loadings.
  - The averaged *variance span* will be close to 100%.
2. Models can be built overestimating the number of components. They will present:
  - Mid/Low *variance span* for the components that exceed the proper number of components (due to overestimation).
  - The averaged *variance span* will be lower the more overestimated components are in the model.
3. Models can be built with an appropriate number of components but they can present wrong shape of loadings. These will show:
  - Low *variance span* per component
  - Low averaged *variance span*.



For MB-PLS, we have seen that one can have a high variance span even for exceeding components. When results from the explained variance criterion are not clear, (namely, when the variance span is neither definitely close to one, nor very low) we suggest to match it with the investigation of the correlation coefficients between the simulated scores/loadings and the scores/loadings estimated from the model. This further inspection gives an additional overview on the relationship between the model parameters and the original ones.

Additionally, MB-PLS presents further model parameters (missing in SO-PLS and PO-PLS), the *super-weights* [8-11], which can be interpreted. They can be investigated to understand the contribution of the different blocks to the prediction of the response; high super-weights' values correspond to high contributions.

### 3. Simulation study

Different simulations were conducted to investigate the behavior of MB-PLS, SO-PLS and PO-PLS in handling blocks of different dimensionality (underlying dimensionality and the number of variables) and handling together categorical (design variable blocks) and non-categorical blocks. Predictions from the three methods will be compared, but the discussion will be more focused on the interpretation of the different models than on predictions. In order to achieve more general results, simulations have been replicated one hundred times. This means that all the simulated model parameters have been re-generated for each one of the one hundred data sets.

All the simulated data sets are divided into *training data set* (constituted by  $\mathbf{X}$ ,  $\mathbf{Z}$  and  $\mathbf{y}$ ) for the construction of the calibration models, and *test data set* (constituted by  $\mathbf{X}_t$ ,  $\mathbf{Z}_t$  and  $\mathbf{y}_t$ ) used only for the validation of the models' predictive ability. The selection of the optimal number of components is based on the training set (evaluating the RMSECVs, as explained in Section 2.4). The simulation study is divided in *Part I* and *Part II*.

#### 3.1 Simulation study Part I

##### *Simulated Data sets*

The  $\mathbf{X}$ -block is constructed to resemble spectroscopic measurements with five hundred variables and eight components ( $\mathbf{C}_X$ ).  $\mathbf{Z}$  mimics process variables, and has only fifteen variables and two components ( $\mathbf{C}_Z$ ). The two blocks are constructed to have one common component  $\mathbf{T}_{C(N \times 1)}$ . Consequently,  $\mathbf{X}$  has one common and seven distinct components ( $\mathbf{T}_{DX(N \times 7)}$ ) and  $\mathbf{Z}$  has the common component and one distinct ( $\mathbf{T}_{DZ(N \times 1)}$ ).  $\mathbf{X}$ -scores  $\mathbf{T}_x$  and  $\mathbf{Z}$ -scores  $\mathbf{T}_z$  are generated as  $\mathbf{T}_x = [\mathbf{T}_C \mathbf{T}_{Dx}]$  and  $\mathbf{T}_z = [\mathbf{T}_C \mathbf{T}_{Dz}]$ .

All the scores are simulated from the normal distribution  $N(0,1)$ .

Both  $\mathbf{X}$  and  $\mathbf{Z}$  are generated as a TP-product, i.e. scores multiplied by loadings. The  $\mathbf{X}$ -loadings,  $\mathbf{P}_X$ , are simulated as sums of Gaussians distributions. The  $\mathbf{Z}$ -loadings,  $\mathbf{P}_Z$ , are simulated in such a way that one or two of the variables per component have a value between 0.75 and 1.25 while all the others have values between -0.25 and 0.25. This means that only a few of the fifteen variables affect the response.

An example of simulated loadings is shown in Figure 1;  $\mathbf{X}$ -loadings are shown in Figure 1(a) and  $\mathbf{Z}$ -loadings for the two components in Figure 1(b).

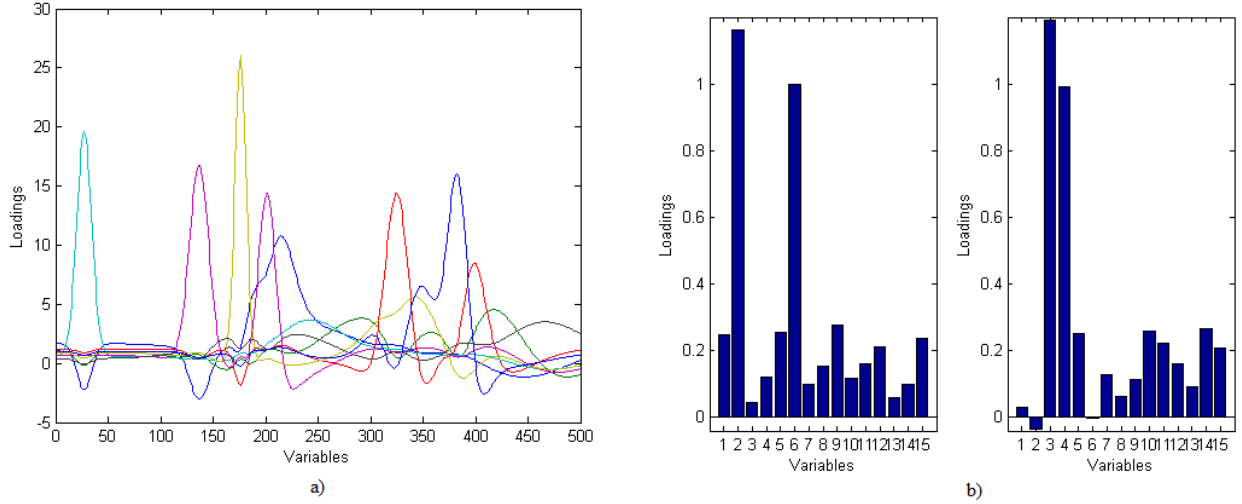


Figure 1 Loadings for the simulation study. a) X-loadings (eight components) b) Z-loadings (two components)

The response,  $\mathbf{y}_{(N \times 1)}$ , is generated as:

$$\mathbf{y} = \mathbf{T}\mathbf{b}$$

Where  $\mathbf{b}_{(10 \times 1)}$  is the coefficient vector generated as a matrix containing random values drawn from the uniform distribution in the open interval (0.05, 1.05).  $\mathbf{T}$  is obtained by:  $\mathbf{T} = [\mathbf{T}_c \mathbf{T}_{Dx} \mathbf{T}_{Dz}]$ . Finally, random noise (corresponding to the 5% of the signal) was added to all the blocks (predictors and response), both in the training and in the test set.

All the blocks in the training set have 70 samples. The test set is simulated in the same way but with 1000 samples.

In order to check the behavior of the methods in handling noise, a further study was conducted simulating one hundred additional (training and test) data sets where the amount of noise added to predictors was 50% (of the signal) and for  $\mathbf{y}$  it was 10%.

### 3.2 Simulation study Part II

The aim is here to investigate the behavior of MB-PLS, SO-PLS and PO-PLS handling a categorical block and a spectroscopic block.  $\mathbf{X}_{des}$  is simulated as a  $2^3$  full factorial design which is repeated nine times (in the row direction), giving a matrix of dimensions  $72 \times 3$ .  $\mathbf{Z}$  and  $\mathbf{y}$  are simulated as explained above, but the number of samples is 72. Blocks in the test set ( $\mathbf{X}_{des_t}$ ,  $\mathbf{Z}_t$  and  $\mathbf{y}_t$ ) are simulated as  $\mathbf{X}_{des}$ ,  $\mathbf{Z}$  and  $\mathbf{y}$  but test samples are 1000. Finally, random noise (correspondent to the 5% of the signal) is added to  $\mathbf{Z}$ ,  $\mathbf{Z}_t$ ,  $\mathbf{y}$  and  $\mathbf{y}_t$ .

Design data can be found in different sample sizes. For the sake of completeness, an additional simulation (equal to the one above exposed, except for the sample size) was conducted repeating the full factorial design matrix two times, ending up with an  $\mathbf{X}_{des}$  of dimension  $16 \times 3$ . Results were comparable to those with 72 samples, therefore they are not reported.

## 4. Results

### 4.1 Simulation study part I-results and discussion

MB-PLS, SO-PLS and PO-PLS models have been constructed using the simulated training set described above and then external validation of predictive ability was carried out by the test set. Prediction

results are reported in Table 1 and the interpretation of the models is given below in the specific sections.

#### 4.1.1 Simulation study part I-Predictions

Firstly, the most appropriate approach (sequential or global) for selecting the number of components in SO-PLS and PO-PLS is investigated. In order to do it, in a preliminary simulation study we compared sequential and global estimation of components and found global to be substantially better than sequential for prediction purposes. We therefore chose to use this throughout the paper.

For the *simulation study Part I*, averaged (over one hundred replicates) RMSEPs from MB-PLS, SO-PLS and PO-PLS are reported in Table 1. MB-PLS predicts slightly better than SO-PLS and PO-PLS. Statistical relevance of the differences has been tested by two way ANOVA (significance level 5%). Then, Tukey's honestly significant difference criterion was used to evaluate which means differ from the others. A graphical representation of results from Tukey's test is reported in Figure 2.

*Table 1 Root Mean Squares Errors for MB-PLS, SO-PLS and PO-PLS in the simulation study Part I (different level of added noise) and Part II*

Simulation study Part I							
Added Noise (%)		RMSEP MB-PLS (averaged over 100 replicates)	Std. Deviation	RMSEP SO-PLS (averaged over 100 replicates)	Std. Deviation	RMSEP PO-PLS (averaged over 100 replicates)	Std. Deviation
X and Z	Y						
5	5	0.37	0.04	0.36	0.03	0.38	0.06
50	10	1.58	0.25	1.86	0.24	2.14	0.38
Simulation study Part II							
5	5	0.37	0.04	0.34	0.03	0.38	0.04

The same simulation has been repeated adding 50% (of the signal) noise to *X*- and *Z*-block and 10% to *y*, in order to check how methods handle very noisy data sets. With such a high noise level, there is a clear difference between the models: MB-PLS has the lowest RMSEP and PO-PLS has the highest.

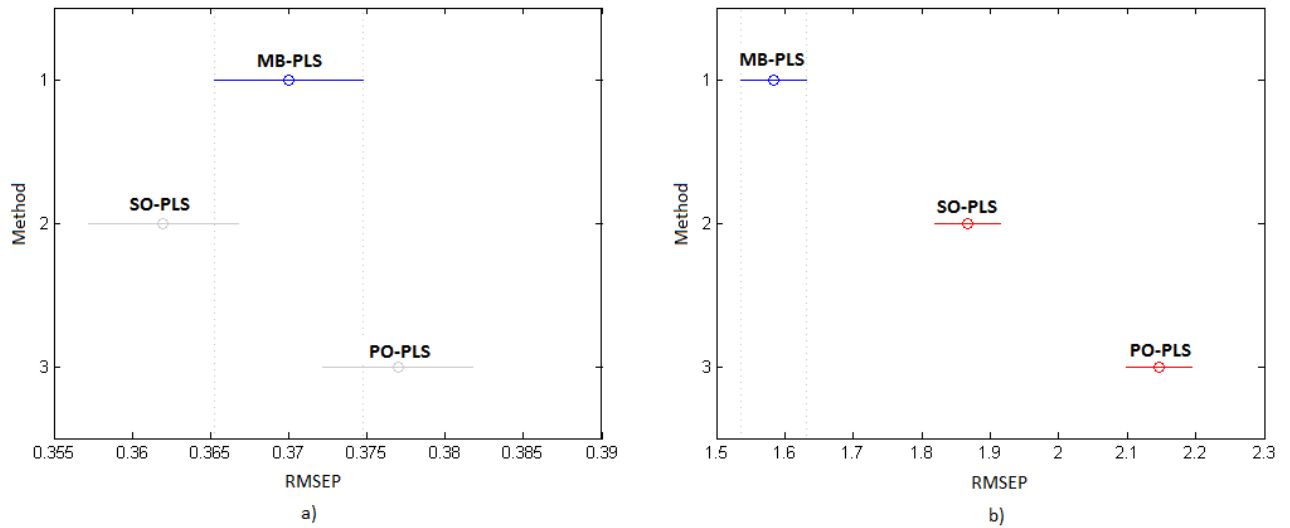


Figure 2 Graphical representation of Tukey's test. a) results after the addition of 5% noise on both predictors and responses. b) results after the addition of 50% noise to predictors and 10% to responses.

In sequential methods such as SO- and PO-PLS, the order of the blocks could affect the models. The same simulation has been repeated inverting the order of the blocks. This means that the  $\mathbf{X}$ -block is the process-like one, while  $\mathbf{Z}$  is the spectra-like. Results (not shown) are in agreement with those in Table 1; confirming that the order of the blocks is not affecting predictions.

#### 4.1.2 Simulation study part I-Interpretation

Interpretation is here focused on models where the added noise is 5% of the signal. As explained in Section 4.1, the  $\mathbf{X}$ -block is simulated using eight components (one common component between the blocks and seven distinct components). The  $\mathbf{Z}$ -block is based on the common component and the distinct one. Since the noise is not that high, the number of components selected in each model is expected to be close to these numbers.

Number/combinations of components selected in the one hundred models are shown in Figure 3

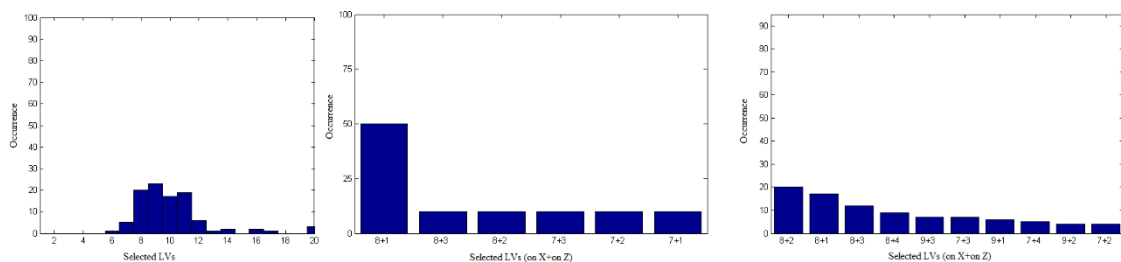


Figure 3: Histograms showing the number of components selected in the one hundred MB-PLS, SO-PLS and PO-PLS models. Left subplot: selected components in MB-PLS models. Central subplot: selected components in SO-PLS models. Right subplot: selected components in PO-PLS models. One common component is selected in each repetition (not shown).

In MB-PLS, the expected number of selected components is nine (one common component plus seven distinct in  $\mathbf{X}$  and one distinct in  $\mathbf{Z}$ ). In Figure 3 is shown that MB-PLS selects mainly from eight to eleven components. In SO-PLS, the optimal complexity corresponds to eight plus one or seven plus two components (depending on in which block the model extracts the common component). Eight plus one occurs in 50% of models and seven plus two components are chosen ten times, showing that the common component is mainly extracted from the  $\mathbf{X}$ -block. In conclusion, SO-PLS

selects the expected number of components in 60% of models. In the other models, components are overestimated (30%). Only in ten models, the total components are less than nine. Concerning PO-PLS, the expected number of components is one (common) plus seven (distinct) for the **X**-block plus one (distinct) for the **Z**-block. One common component has been extracted one hundred times (not shown). Instead, the combination of seven plus one distinct components has never been used. Often, variables in both blocks are overestimated.

The explained variance criterion has been applied on a model built on a simulated data set with 5% noise. Results are discussed below and summarized in Table 2.

Simulated and estimated loadings appeared comparable, confirming that the interpretation of these loadings would lead to the interpretation of actual information present in the blocks.

In the example reported, in order to show the potential of this kind of investigation, more component than necessary have been extracted from the model. Looking at the variance span, one can have an indication on which (and how many) loadings resemble the originals.

*Table 2 Variance span from the prediction of loadings extracted from MB-PLS, SO-PLS and PO-PLS model by OPLS on the original loadings. In the last row is reported the actual number (by construction) of components for the blocks.*

Components	MB-PLS		SO-PLS		PO-PLS			
	<b>X</b>	<b>Z</b>	<b>X</b>	<b>Z</b>	Distinct <b>X</b>	Distinct <b>Z</b>	Common <b>X</b>	Common <b>Z</b>
1	0.97	1.00	1.00	0.98	1.00	0.98	1.00	1.00
2	0.94	1.00	1.00	0.89	1.00	0.04		
3	0.95	1.00	1.00	0.38	1.00	0.27		
4	0.94	1.00	1.00	0.05	1.00	0.63		
5	0.78	0.99	1.00	0.03	1.00	0.01		
6	0.95	0.99	1.00		1.00			
7	0.90	1.00	1.00		1.00			
8	0.79	1.00	1.00		0.96			
9	0.82	0.82	0.98		0.00			
10	0.90	0.02	0.00		0.00			
11	0.62	0.18	0.00		0.00			
12	0.56	0.28	0.00		0.00			
# of expected LVs	8+2		7 8	2 1	7	1	1	1

This approach reveals the actual complexity required by each method. In SO-PLS, the variance span drop after nine components in **X** and two in **Z**; this suggest the SO-PLS model has one interpretable component more than expected per block. For PO-PLS, the explained variance criterion indicates to interpret eight and one distinct components plus the common; so one component (on **X**) more than expected is interpretable. In MB-PLS variance span are slightly fluctuant, probably due to the combined contribution of the two blocks to the model. Variance span relevantly drop after **Z**-loading nine, but they only decrease after **X**-loading ten. In order to achieve a more clear conclusion, further inspections of MB-PLS' loadings and super-scores is reported below.

The overestimation of components presented can be due to the fact that models may need additional components to handle noise.

In Figures 4, 6 and 7 loadings extracted from MB-, SO- and PO-PLS models respectively, are shown.

Looking at  $X$ -loadings in Figure 4 (MB-PLS), it is quite evident that after loading vector seven, noise starts being modelled. The  $X$ -block contributes to all the loading vectors, instead, the  $Z$ -block contributes mainly to loading vectors one, two, three and twelve (i.e. four components in total). This is also confirmed from the super-weights plot reported in Figure 5.

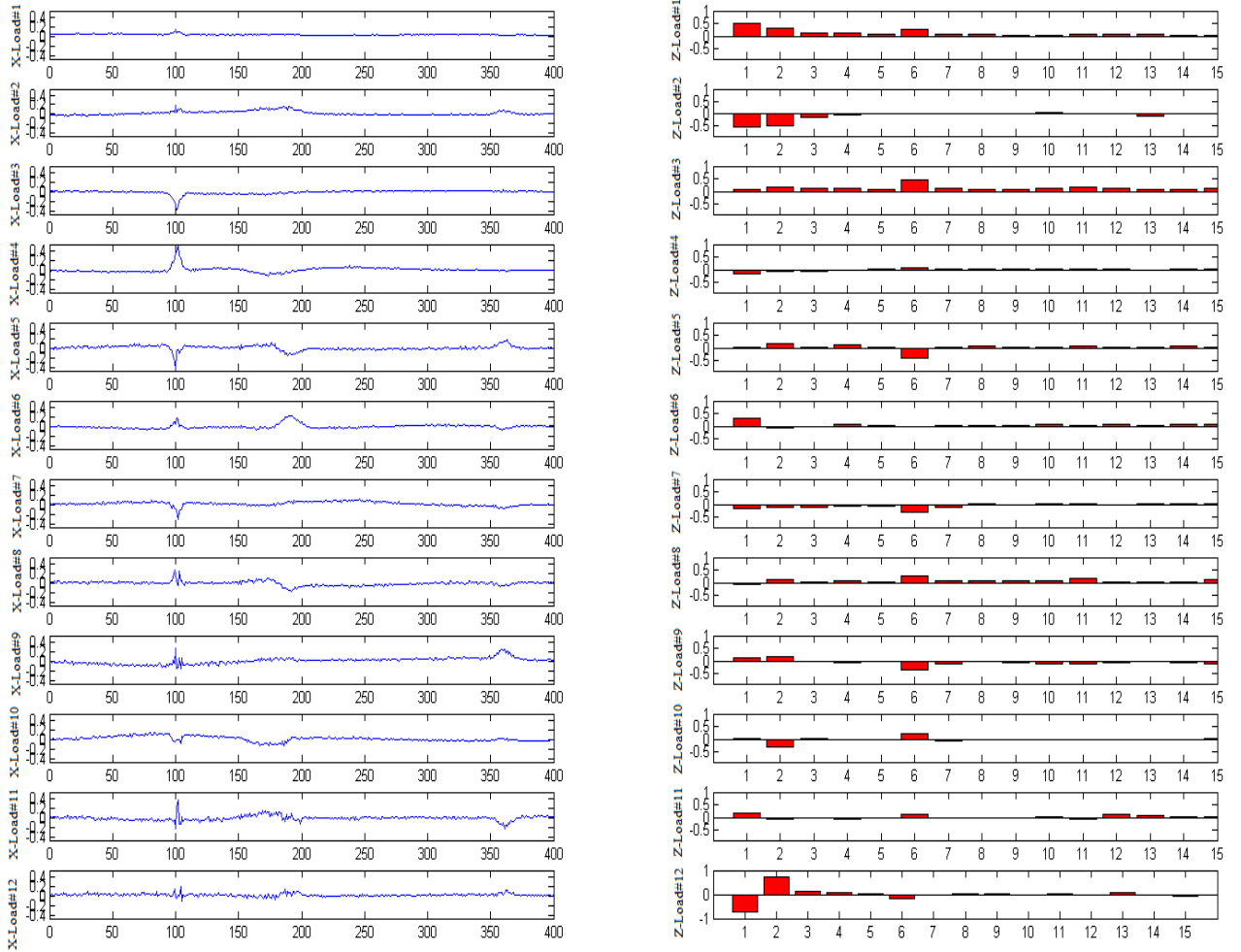


Figure 4: Loadings extracted from the MB-PLS model.

It is clear from the super-weights that  $X$  contributes to all components, while  $Z$  contributes mainly to components one, two, three, and five. Consequently, components from six to eleven plus component four are constituted mainly by  $X$ , while the other components get contributions from both blocks.

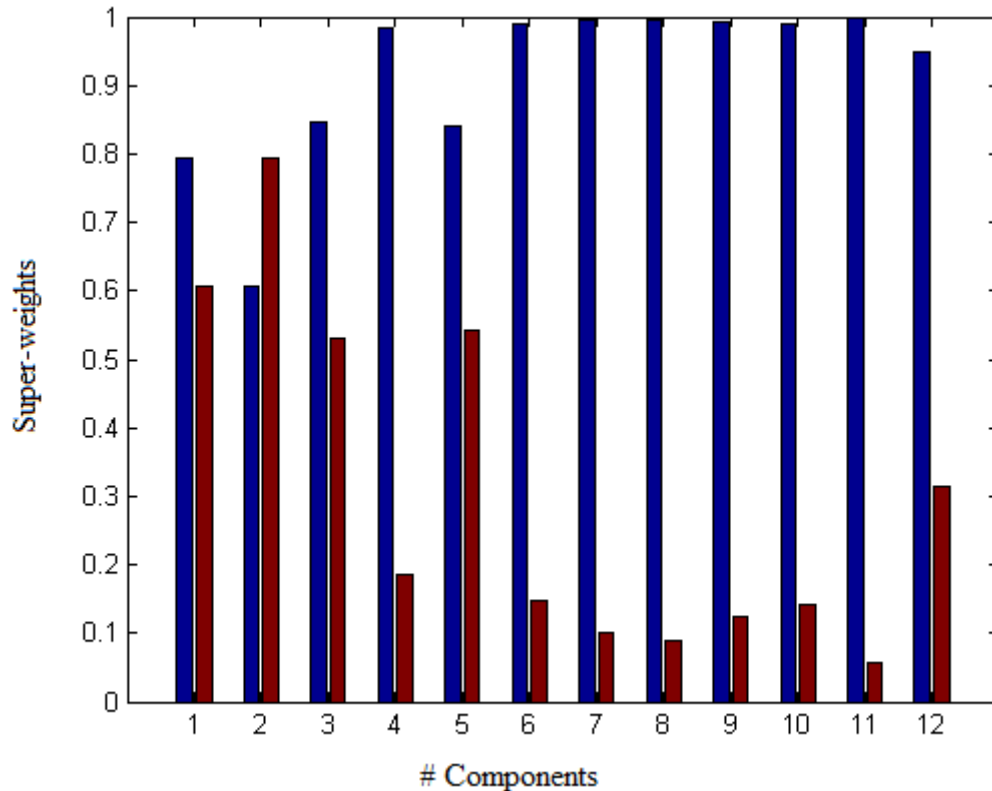


Figure 5: Super-weights: Contribution from the  $X$ -block in blue, from the  $Z$ -block in red.

A further tool to inspect the number of interpretable components is checking the correlation between simulated scores and scores estimated from the MB-PLS model. Inspecting these correlations it comes out that estimated scores eleven and twelve have low correlation with simulated  $X$ -scores, and  $X$ -score ten has medium correlation with simulated  $X$ -scores. Concerning  $Z$ -scores, only those higher than three have low correlation with simulated  $Z$ -scores. Looking at this together with the explained variance criterion, it appears that nine  $X$ -components and three  $Z$ -components could be interpreted. These results are in a quite good agreement with the expected actual complexity of the system.

Concerning the interpretation of the SO-PLS models, one should take into account that one important aspect of the SO-PLS method is the possibility of interpreting the additional contribution added from the  $Z$ -block to the model. Investigating SOPLS'  $Z$ -loadings, one gets an overview of information present in  $Z$  but not in  $X$ .

In Figure 6 are shown loadings from the SO-PLS model.  $X$ -loadings are in blue on the left side of the plot, while  $Z$ -loadings are on the right side in red.

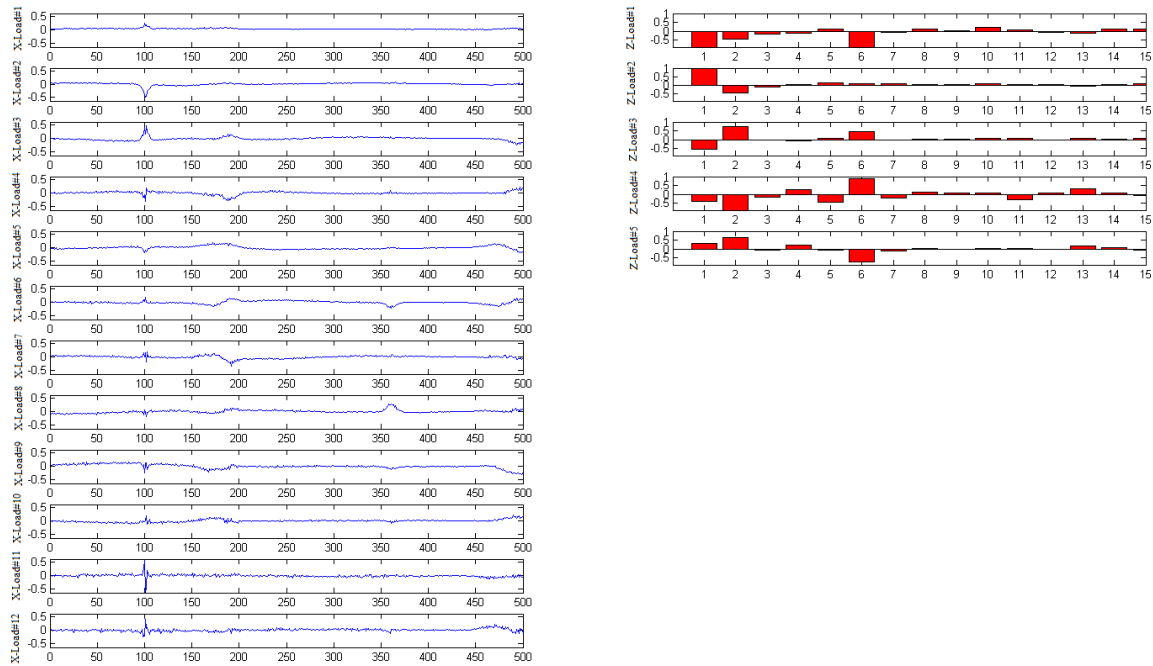


Figure 6: Loadings extracted from the SO-PLS model.  $X$ -loadings in blue, left side.  $Z$ -loadings in red, right side.

Both blocks seem to contribute to the components, but the interpretation of  $X$ -components higher than nine and  $Z$ -components higher than two would be misleading. It is possible to recognize noise from  $X$ -loading nine to  $X$ -loading twelve. Looking at  $Z$ -loadings (Figure 6, red bars), one could (erroneously) interpret all of them. This highlights the relevance of the choice of a proper number of components to be used building the calibration model.

In Figure 7 loadings extracted from the PO-PLS model are shown.  $X$ -components are on the left side, while  $Z$ -components are on the right one. The first subplot on both sides represent the common components (extracted from  $X$  and  $Z$ , respectively). The conclusions regarding the interpretation of these loadings are similar to those from SO-PLS. Namely, components seem constituted from both blocks. It is possible to recognize noise in  $X$ -components higher than seven; it is not straightforward to see noise in  $Z$ -components higher than one. Additionally to the relevance of the determination of the optimal complexity for calibration models, this indicates that avoiding the interpretation of noise in process variables is intrinsically trickier than avoiding it interpreting instrumental signals with a pattern.



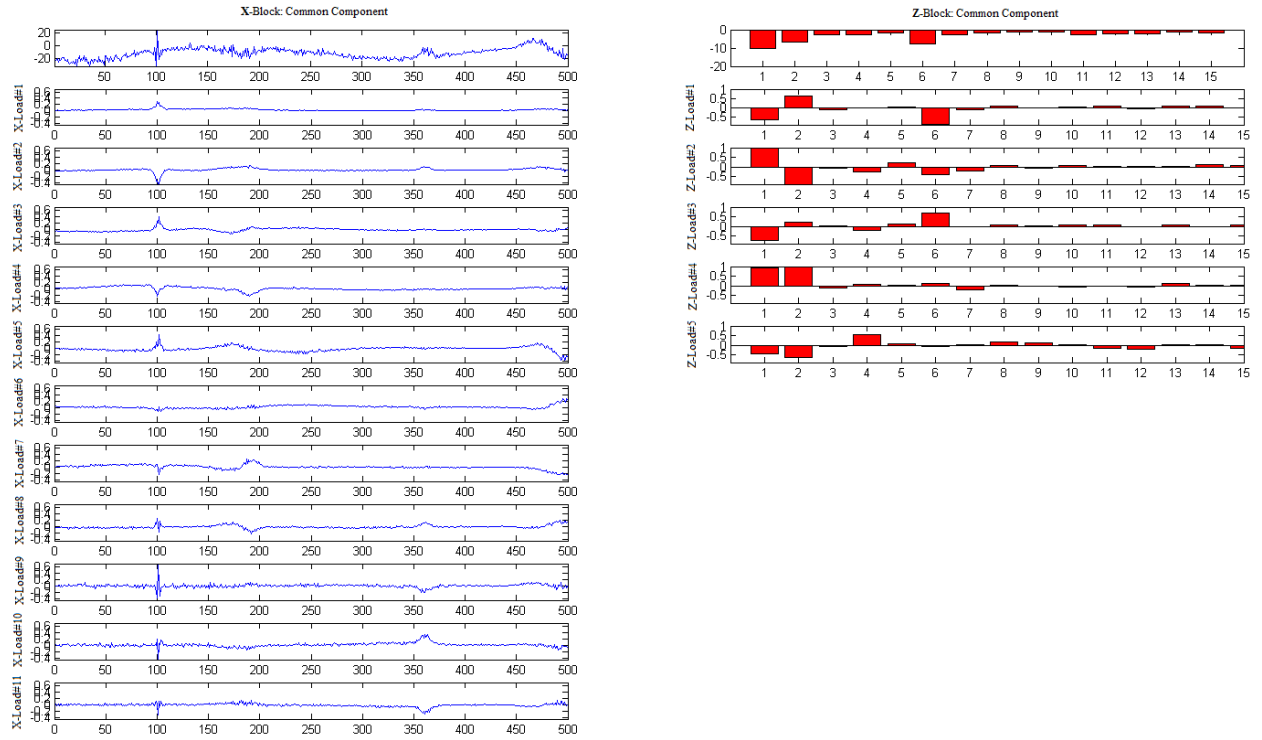


Figure 7: Loadings extracted from the PO-PLS model.  $X$ -loadings in blue, left side.  $Z$ -loadings in red, right side

## 4.2 Simulation study part II-results and discussion

MB-PLS has been constructed as exposed in Section 2.1. For SO-PLS and PO-PLS, the number of components for the  $X$ -block has been fixed to three. In this way, the regression on  $X$  is identical to ordinary least squares. Then, external validation was carried out by the test set.

Since the  $X$ -block is a categorical block and SO-PLS and PO-PLS models have been forced to select three components the interpretation of  $X$ -loadings is not discussed (and  $X$ -loadings plots are not shown).

### 4.2.1 Simulation study part II-Predictions

Averaged (over the one hundred replicates) RMSEPs from MB-PLS, SO-PLS and PO-PLS are reported in Table 1. In table, the difference between SO-PLS and the other two methods is appreciable. In fact, MB-PLS and PO-PLS give comparable results, definitely worse than SO-PLS. The differences between RMSEPs have been tested by two way ANOVA and Tuckey's test (see Section 4.1.1 for more details). Test's results are graphically reported in Figure 8. SO-PLS handles the categorical block better than the other two methods. Considering MB-PLS, this is not completely surprising. In fact, the presence of categorical data could lead to an overestimation of the components needed from the model, affecting negatively predictions [20].

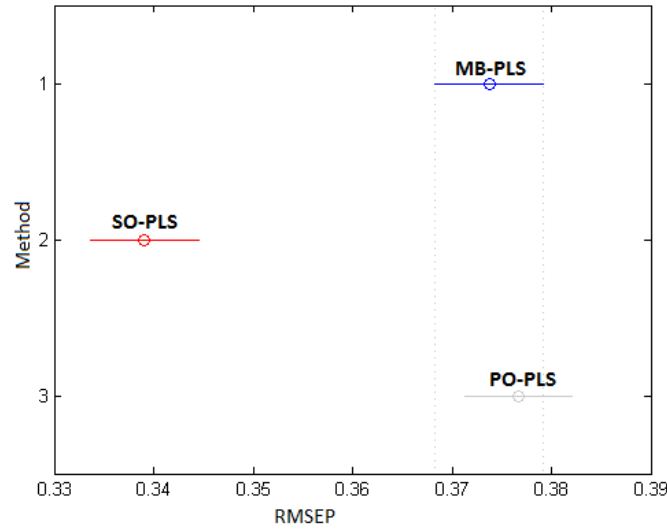


Figure 8: Graphical representation of Tukey's test.

#### 4.2.2 Simulation study part II-Interpretation

As explained above, the number of components required by the regression involving the  $\mathbf{X}$ -block in SO- and PO-PLS was fixed to three. Components for MB-PLS, and for  $\mathbf{Z}$ -block in SO-PLS and PO-PLS are defined as explained in Section 2.4. In Figure 9 selected number of components and their occurrence over the one hundred replicates are shown. In PO-PLS, only one common component has always been selected, therefore it is not reported in the figure.

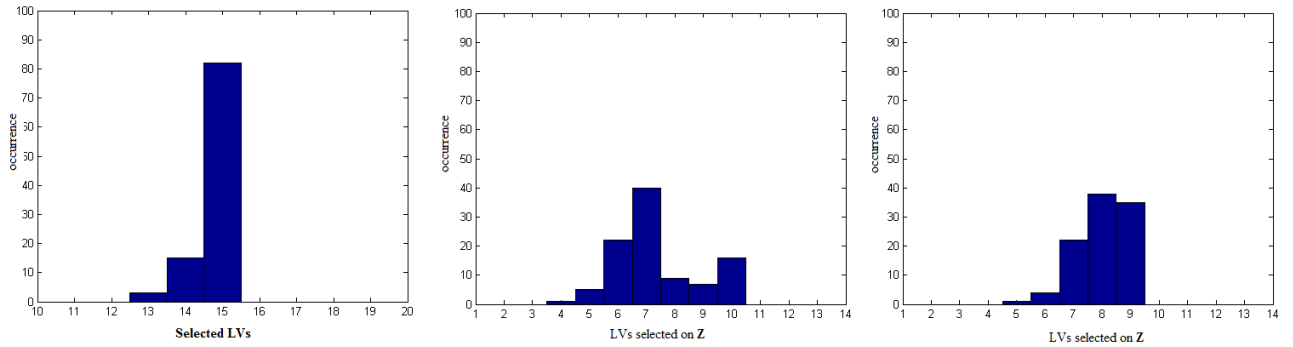


Figure 9: Histograms showing the number of components selected in the one hundred MB-PLS, SO-PLS and PO-PLS models. Left subplot: selected components in MB-PLS models. Central subplot: selected components from  $\mathbf{Z}$  in SO-PLS models. Right subplot: selected distinct  $\mathbf{Z}$ -components in PO-PLS models

According to the simulation, the expected number of components is ten (three for  $\mathbf{X}$  plus seven for  $\mathbf{Z}$ ) for MB-PLS, and the expected number of  $\mathbf{Z}$ -components is seven for SO-PLS. For PO-PLS, we expect one common component plus (two distinct for the  $\mathbf{X}$ -block and) seven distinct for the  $\mathbf{Z}$ -block. In Figure 9 is evident that MB-PLS is always overestimating the number of components. In SO-PLS, the most frequent selected number of components is the expected one (seven components are required in 40% of the models). Otherwise, six or ten components are selected (in twenty and fifteen models over one hundred, respectively). In PO-PLS, mainly eight or nine components are selected. Overestimation of components is an expected drawback of applying PLS regression on categorical data [20].

In order to inspect models built combining categorical plus spectra-like blocks, one of the one hundred data sets has been deeply investigated and described below.

MB-PLS model required nine components. In SO-PLS, seven components were selected on  $\mathbf{Z}$ . PO-PLS has one common component and eight distinct components for the  $\mathbf{Z}$ -block. The similarity between the original loadings and those extracted from the models has been tested as explained in 2.5. As can be seen in Table 3, all the extracted loadings result interpretable. Loadings are reported in Figures 10, 12 and 13.

Table 3 Variance span from the prediction of loadings extracted from MB-PLS, SO-PLS and PO-PLS model by OPLS on the original loadings.

	MB-PLS	SO-PLS	PO-PLS	
LVs	$\mathbf{Z}$	$\mathbf{Z}$	Distinct $\mathbf{Z}$	Common $\mathbf{Z}$
1	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	
3	1.00	1.00	1.00	
4	1.00	1.00	1.00	
5	1.00	1.00	1.00	
6	1.00	1.00	1.00	
7	1.00	1.00	0.99	
8	1.00		1.00	
9	0.99			

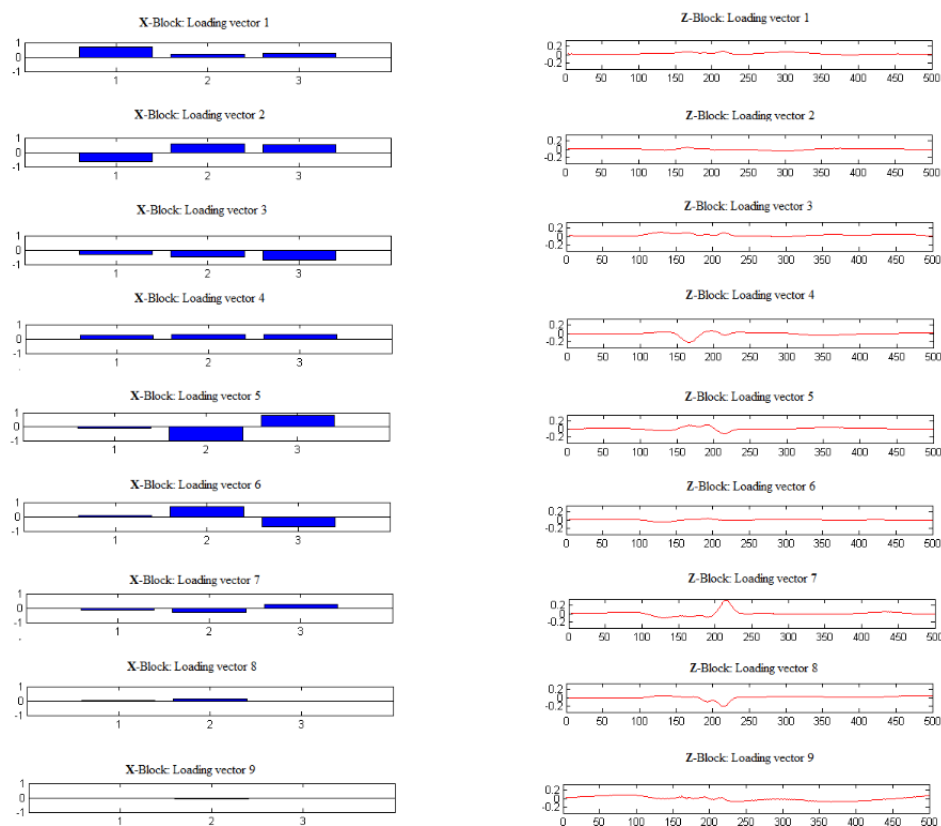


Figure 10: Loadings extracted from the MB-PLS model.

In Figure 10 are reported loadings extracted from the MB-PLS model. The categorical block mainly contributes to loadings one, two, five and six (and to a lesser extent, three).  $\mathbf{Z}$  contributes (with

different emphasis) to all the components. Components four, seven, eight and nine consist mainly of spectral variables, while  $\mathbf{Z}$ 's contribution to components two and six is weak. This interpretation is confirmed also looking at the super-weights in Figure 11.

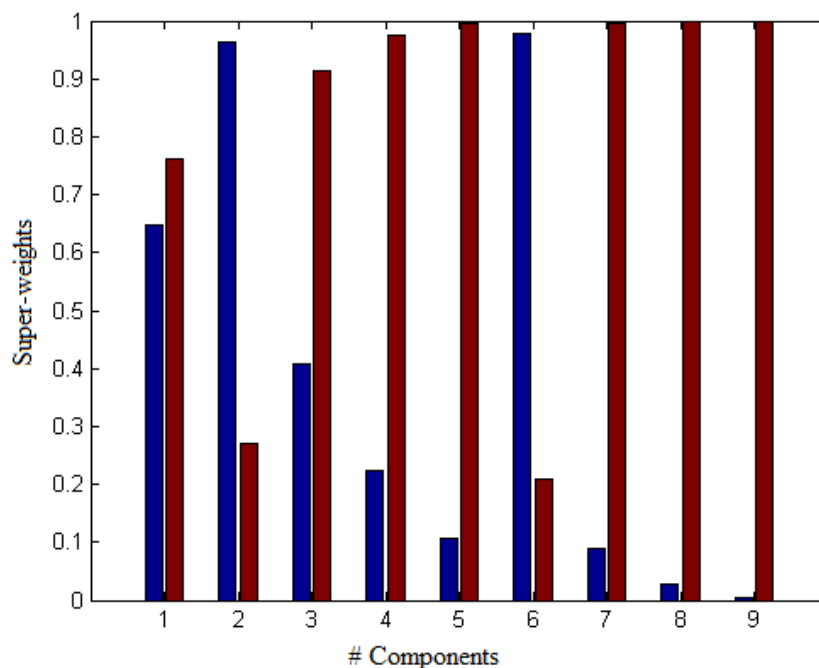


Figure 11: Super-weights: Contribution from the X-block in blue, from the Z-block in red.

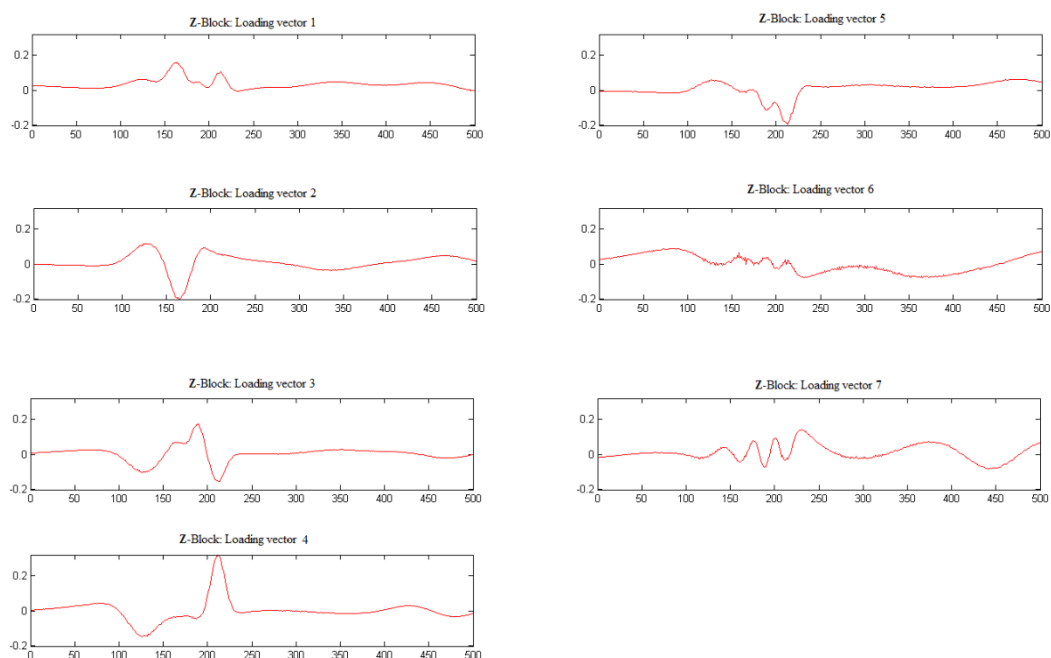


Figure 12:  $\mathbf{Z}$ -Loadings extracted from the SO-PLS model.

In Figure 12 are shown loadings from the SO-PLS model.  $\mathbf{Z}$ -loadings show clearly their shape, and would be easily interpreted. Only loading vector six present a bit of noise, but not enough to enable the interpretation.

In Figure 13 the  $\mathbf{Z}$ -loadings from the PO-PLS model are reported.

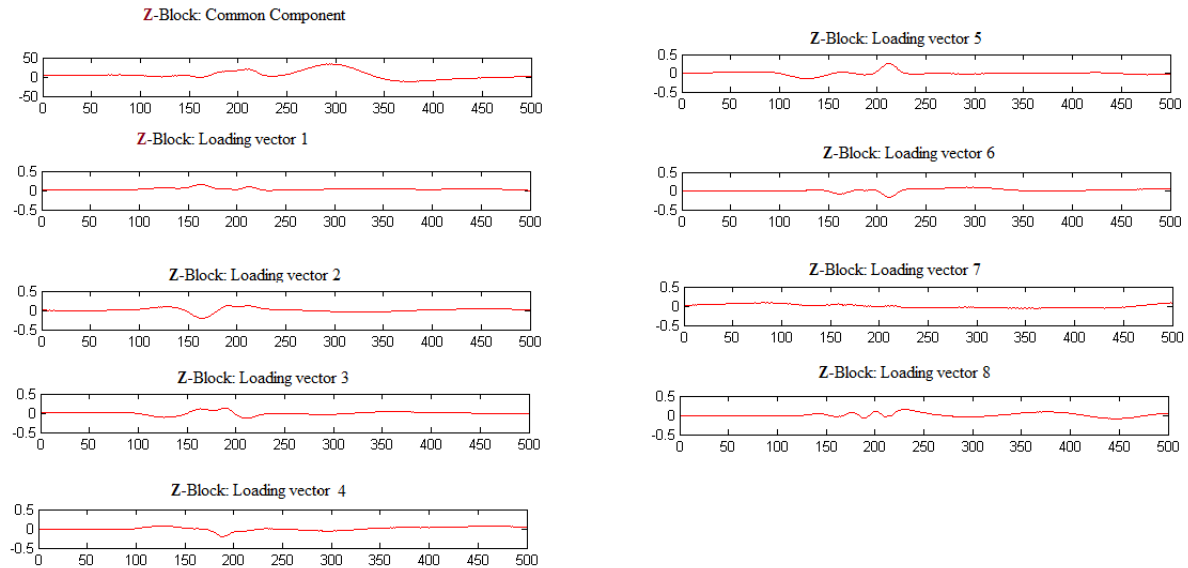


Figure 13: Loadings extracted from the PO-PLS model

The first subplots on the left side in Figure 13 show the common component in the  $\mathbf{Z}$ -block, all the others display distinct ones. The spectra-like block is visibly contributing to all the distinct components, except for loading seven (and to a lesser extent, loading one).

## 5. Discussion and Conclusions

### 5.1 Combination of process and spectra variables

MB-PLS gives slightly better predictions than SO-PLS and PO-PLS, especially when the data is very noisy.

In MB-PLS, the number of components cannot be fixed separately for each block, which leads to a natural overestimation of components. This makes MB-PLS the most complicated method (among the three) to interpret. In the data set presented in Section 4.1.2, the actual complexity is nine. The explained variance criterion did not give a straight indication about the number of interpretable components. The investigation of the super-weights (Figure 5) and the inspection of the correlation among scores reveal that no more than nine and three components could be interpreted for  $\mathbf{X}$  and  $\mathbf{Z}$ , respectively. This indicates that a slight overestimation is required by MB-PLS (probably to handle noise).

SO-PLS generally requires a number of components in agreement with the actual complexity of the blocks or slightly higher. This overestimation is not huge, and probably due to noise.

Over the one hundred replicates, PO-PLS overestimates the number of components required, but less than MB-PLS.

A comparison between the actual and the required complexity among the three methods is reported in Table 4.

Table 4. Actual and required complexity among MB-PLS, SO-PLS and PO-PLS.

	MB-PLS	SO-PLS ( $X + Z$ )	PO-PLS (Com. + $X_{\text{dist.}}$ + $Z_{\text{dist.}}$ )
Actual complexity	9	$7+2/8+1$	$1+7+1$
Interpretable	9	$9+3$	$1+8+1$

## 5.2 Combination of categorical blocks and spectra variables

In the simulation study, SO-PLS gives better predictions than MB-PLS and PO-PLS.

From the study appears that SO-PLS often requires a number of components close to the expected one. Instead, PO-PLS and (even more) MB-PLS overestimate the number of components required; which may lead to overfitted models.

Loadings extracted from the different models result interpretable for all the methods. SO-PLS shows that it is modelling a bit of noise, but its entity is not enough to enable the interpretation.

## 5.3 Conclusions

From the interpretation point of view, the same conclusions are reached in the different simulation studies: for investigation of model parameters such as loadings, MB-PLS is the less preferable method among the three. This is due to the fact that in MB-PLS, components cannot be chosen independently for  $X$  and  $Z$ . As a consequence, the number of components selected is not the most appropriate for each block, and the interpretation of the model parameters could be misleading/too complex. SO- and PO-PLS give interpretable models, the main issue is the definition of the optimal complexity, in particular for PO-PLS, which is much more prone (than SO-PLS) to overestimate components.

## 6. References

- [1] R. Bro, F. van den Berg, A. Thybo, C.M. Andersen, B.M. Jørgensen, H. Andersen, Multivariate data analysis as a tool in advanced quality monitoring in the food production chain, Trends Food Sci. Technol. 13 (2002) 235–244.
- [2] S. Hassani, H. Martens, E.M. Qannari, M. Hanafi, G.I. Borge, A. Kohl, Analysis of –omics data: graphical interpretation- and validation tools in multi-block methods, Chemometr. Intell. Lab. Syst. 104 (2010) 140–153.
- [3] C. Zhao, F. Gao, Multiblock-based qualitative and quantitative spectral calibration analysis, Ind. Eng. Chem. Res. 49 (2010) 8694–8704.
- [4] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, J. Chemometr. 15 (2001) Pages 715–742.
- [5] El Ghaziri, V. Cariou, D.R. Rutledge, M. Qannari, Analysis of multiblock datasets using ComDim: Overview and extension to the analysis of  $(K + 1)$  datasets, J. Chemometr. 30 (2016) 420–429.
- [6] S. Bougeard, M. Qannari, N. Rose, Multiblock redundancy analysis: Interpretation tools and application in epidemiology, J. Chemometr. 25 (2011) 467–475.
- [7] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometr. 10 (1996) 463–482.
- [8] I. E. Frank, J. Feikema, N. Constantine and B. R. Kowalski, J. Chem. Info. Comput. Sci. 24 (1984) 20–24.
- [9] I. E. Frank and B. R. Kowalski, Prediction of wine quality and geographic origin from chemical measurements by Partial Least-Squares regression modeling, Anal. Chim. Acta, 162 (1984) 241–251.

- [10] L. E. Wangen and B. R. Kowalski, A multiblock Partial Least Squares algorithm for investigating complex chemical systems *J. Chemometrics* 3 (1988) 3–20.
- [11] J.A. Westerius, T. Kourti, J.F. MacGregor, Analysis of hierarchical PCA and PLS models, *J. Chemometr.* 12 (1998) 301–321.
- [12] T. Næs, O. Tomic, B.-H. Mevik, H. Martens, Path modelling by sequential PLS regression, *J. Chemometrics* 25 (2011) 28–40.
- [13] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis, *Chemometr. Intell. Lab. Syst.* 124 (2013) 32–42.
- [14] Ingrid Måge, Bjørn-Helge Mevik, Tormod Næs, Regression models with process variables and parallel blocks of raw material measurements, *J. Chemometrics* 22 (2008) 443–456.
- [15] I. Måge, E. Menichelli, T. Næs, Preference mapping by PO-PLS: separating common and unique information in several data blocks, *Food Qual. Prefer.* 24 (2012) 8–16.
- [16] H. Hotelling. Relations between two sets of variates, *Biometrika*, 28 (1936) 321– 377.
- [17] J.R. Kettenring, Canonical analysis of several sets of variables, *Biometrika* 58 (1971) 433–451.
- [18] U. Indahl, A twist to partial least squares regression, *J. Chemometr.* 19 (2005) 32–44.
  
- [19] A. Biancolillo, I. Måge, T. Næs, Combining SO-PLS and linear discriminant analysis for multi-block classification, *Chemometr. Intell. Lab. Syst.* 141 (2015) 58–67.
- [20] K. Jørgensen, T. Næs, A design and analysis strategy for situations with uncontrolled raw material variation, *J. Chemometrics* 18 (2004) 45–52.