# Variable selection in multi-block regression

Alessandra Biancolillo [a,b,*], Kristian Hovde Liland[a,c], Ingrid Måge[a], Tormod Næs[a,b], Rasmus Bro[b]

[a] *Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway*

[b] *Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[c] *Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N-1432 Ås, Norway*

*Corresponding author:

Tel: +47 64 97 01 15

e-mail: alessandra.biancolillo@nofima.no

## 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.

## Keywords

Variable Selection; Multi-block; SO-PLS; MB-PLS; Raman; Sensory.

## 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 $\boldsymbol{Y}$-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:

$$\boldsymbol{Y} = \boldsymbol{X}f + \boldsymbol{Z}g + \boldsymbol{E} \qquad (1)$$

Where: $\boldsymbol{X}$ $(N \times J)$ and $\boldsymbol{Z}$ $(N \times L)$ are the predictor blocks and $\boldsymbol{Y}$ $(N \times 1)$ is the response variable. $\boldsymbol{E}$ $(N \times 1)$ is the residual matrix and $f$ and $g$ are 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 $\boldsymbol{X}_{conc}$. 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. $\boldsymbol{Y}$ is fitted to $\boldsymbol{X}$ by PLS-regression
2. $\boldsymbol{Z}$ is orthogonalized (obtaining $\boldsymbol{Z}_{orth}$) with respect to the scores of the previous PLS model
3. $\boldsymbol{Y}$ residuals from the first PLS are fitted to $\boldsymbol{Z}_{orth}$
4. The full predictive model is computed by summing up the two contributions from $\boldsymbol{X}$ and $\boldsymbol{Z}$.

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 $\boldsymbol{X}$-block is interpreted by inspecting the PLS model in step 1. The interpretation of the $\boldsymbol{Z}$-block is best done by calculating loadings by projecting $\boldsymbol{Z}$ 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ after variable selection as $\boldsymbol{X_{Red}}$ and $\boldsymbol{Z_{Red}}$.

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

| Variable Selection Method | Multiblock Method | |
|---|---|---|
| | MB-PLS | SO-PLS |
| VIP | ✓ | ✓ |
| SR | ✓ | ✓ |
| Forward Selection | | ✓ |

### 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 Paragraph 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 $\boldsymbol{X_{Conc}}$. Variable selection is then based on the obtained PLS model. This leads to a number of variables being selected and $\boldsymbol{X_{Conc}}$ is reduced obtaining $\boldsymbol{X_{Red}}$. Finally, a new calibration model is obtained for $\boldsymbol{Y}$ using the reduced matrix $\boldsymbol{X_{Red}}$ 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 Paragraph 3.1). In other words, $\boldsymbol{Y}$ is fitted to $\boldsymbol{X}$ and to $\boldsymbol{Z}$ 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, $\boldsymbol{X_{Red}}$ and $\boldsymbol{Z_{Red}}$ are obtained and used in the SO-PLS regression. Compared to the procedure in 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 Paragraph 3.1). Due to the sequential nature of the SO-PLS method, variables can be selected (by VIP or SR) from the $\boldsymbol{X}$-block, from the $\boldsymbol{Z_{Orth}}$-block or from both. When the variable selection involves both blocks, the algorithm is the following:

1. $\boldsymbol{Y}$ is fitted to $\boldsymbol{X}$ by a PLS model.
2. A variable selection method is applied to $\boldsymbol{X}$ obtaining $\boldsymbol{X_{Red}}$.
3. $\boldsymbol{Y}$ is refitted to $\boldsymbol{X_{Red}}$.
4. $\boldsymbol{Z}$ is orthogonalized with respect to the scores of the PLS model in step 3.
5. The residual matrix from step 3. is fitted to $\boldsymbol{Z_{Orth}}$.
6. A variable selection method is applied to $\boldsymbol{Z_{Orth}}$ obtaining $\boldsymbol{Z_{Orth,Red}}$.
7. A new PLS regression is carried out using the reduced matrix $\boldsymbol{Z_{Orth,Red}}$ 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 $\boldsymbol{X}$-

block is reduced, the model will coincide with the one built from SO-PLS using $\boldsymbol{X}_{\boldsymbol{Red}}$ and $\boldsymbol{Z}$ 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.

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 Paragraph 3.1).

First, the best predictor is selected from either $\boldsymbol{X}$ or $\boldsymbol{Z}$ based on RMSECV. Next, each of the successively added variables will come either from $\boldsymbol{X}$ or $\boldsymbol{Z}$. The algorithm will test all the possible combinations which result from either adding one variable from the $\boldsymbol{X}$-block keeping the $\boldsymbol{Z}$-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 $\boldsymbol{X}$- and the $\boldsymbol{Z}$-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 $\boldsymbol{X}$-variable. Likewise, $L$-$v_2$ SO-PLS models are built adding one further $\boldsymbol{Z}$-variable to $\boldsymbol{Z}_{\boldsymbol{Red}}$ (retaining only the previously selected $v_1$ predictors in the $\boldsymbol{X}$-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 Paragraph 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 $\boldsymbol{X}$- and $\boldsymbol{Z}$-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.

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

| | X-Block | | | Z-Block | | |
|---|---|---|---|---|---|---|
| ***Simulation*** | ***# Selective variables*** | ***# Relevant but non-*** | ***# Irrelevant variables*** | ***# Selective variables*** | ***# Relevant but non-*** | ***# Irrelevant variables*** |

|  |  | *selective variables* |  |  | *selective variables* |  |
|---|---|---|---|---|---|---|
| Sim1 | 30 | 30 | 30 | 40 | 40 | 40 |
| Sim2 | 50 | 20 | 20 | 60 | 20 | 40 |
| Sim3 | 80 | 20 | 20 | 60 | 30 | 0 |
| Sim4 | 120 | 0 | 0 | 100 | 0 | 30 |
| Sim5 | 350 | 100 | 50 | 300 | 100 | 50 |
| Sim6 | 350 | 100 | 100 | 300 | 100 | 0 |

For all the data sets, the $\boldsymbol{X}$-block is simulated by multiplying randomly generated scores ($\boldsymbol{T_X}$) and loadings ($\boldsymbol{P_X}$). Both scores ($\boldsymbol{T_X}$) and loadings ($\boldsymbol{P_X}$) are simulated from the normal distribution *N*(0,1). The $\boldsymbol{T_X}$ has fixed dimensionality $(200 \times 4)$ where only the first three components are 'selective'. $\boldsymbol{P_X}$ 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 $\boldsymbol{X}$-block here is generated following the same procedure used for the generation of $\boldsymbol{X}$ in the simulation presented in Appendix A).

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

$$\boldsymbol{T_{Zsel}} = [\boldsymbol{T_{Z1}}\ \boldsymbol{T_{Z2}}] \qquad (2)$$

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

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

$$\boldsymbol{Y} = [\boldsymbol{T_{Xsel}}\ \boldsymbol{T_{Zsel}}] * \boldsymbol{\beta} \qquad (3)$$

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 $\boldsymbol{Z}$-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 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)

***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 smell data are used in the following analysis as the $\boldsymbol{X}$-block, and the taste as $\boldsymbol{Z}$-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 $^{-1}$), and is the one used as the $\boldsymbol{X}$-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$^{-1}$ and is used as the $\boldsymbol{Z}$-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

## 5. Results

All the proposed procedures discussed in Paragraph 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 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 Paragraph 3.1). In procedure 6, it selects few ‘irrelevant’ variables from the $\mathbf{X}$-block and none from the $\mathbf{Z}$-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 $\mathbf{X}$-block, but many from the $\mathbf{Z}$-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 $\boldsymbol{X}$-block, but more than VIP when it comes to the $\boldsymbol{Z}$. 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 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 $\mathbf{X}$ and it selects more ‘relevant but not selective’ variables than VIP from $\mathbf{Z}$. 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 Paragraph 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.

***Table 4 RMSEPs for the prediction of $y$ (from the simulated multiblock datasets Sim1 and Sim 5) 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 X-block (%)* | | | | *Var.Selected in Z-block (%)* | | | | *RMSEP* |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Sim1*** | | | | | | | | | | |
| | | *Sel (%)* | *RelnSel (%)* | *Irr (%)* | *Noise(%)* | *Sel (%)* | *RelnSel (%)* | *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* |
| ***Sim5*** | | | | | | | | | | |
| *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* |

## 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 comparable. Concerning the interpretation, the main difference is that a different number of variables (in particular in the second block) is 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.

***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 | REMSECV | 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$_{(mean)}$ | 1;4;5;6;8; | None | 1 | 0.25 | 54 |
| *Selection on individual block + SO-PLS* | VIP | 1;4;5;6 | 1;4;5;9 | 2,1 | 0.24 | 56 |
| | SR | 1;2;4;5;6;8 | 4 | 1,1 | 0.23 | 60 |
| | SR$_{(mean)}$ | 1;4;8; | 1;2;4;5;6 | 1 | 0.26 | 48 |
| *Selection integrated in SO-PLS* | VIP | 1;4;5;6 | 1;4;10 | 2,1 | 0.24 | 55 |
| | SR | 1;2;4;5;6;8 | 1;2;4;7;8;10;11;13;14 | 1,1 | 0.25 | 53 |
| | SR$_{(mean)}$ | 1;4;8; | 1;7;10;11;13 | 1 | 0.28 | 48 |
| | Forw. Sel. | 2;3;6 | 8 | 1,1 | 0.21 | 66 |

In most models, SR selects more variables than VIP in $\boldsymbol{X}$, but when it comes to $\boldsymbol{Z}$ it depends on the procedure used. Variable selection by SR does not select $\boldsymbol{Z}$-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 Paragraph 3.1), VIP selects more $\boldsymbol{Z}$-variables than SR; when it is implemented in the SO-PLS building (procedures 5-6 from the list in Paragraph 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 $\boldsymbol{Z}$-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 $\boldsymbol{Z}$-block, but on $\boldsymbol{Z}_{orth}$. 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 $\boldsymbol{Z}_{orth}$; 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 $\boldsymbol{Z}$ 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).

For VIP, it is quite consistent in its selection on $\boldsymbol{X}$, independently of the method/model. VIP always selects variables number 1,4,5,6 (*ripe*, *syntetic*, *lactonic* and *sulfuric*, respectively). When applied to MB-PLS it also selects variable number 2, *tropical*. On the $\boldsymbol{Z}$-block the selection is less consistent, but variable number 4 (*syntetic*) 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. Figure 1 shows the selected variables in the two blocks when Both VIP and SR are integrated into the SO-PLS model. Figure 1(a) shows the selected variables by VIP and Figure 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ is explained by $\boldsymbol{X}$ and then removed from the $\boldsymbol{Z}$-Block. Therefore, smell variables that are selected in the $\boldsymbol{X}$-Block are not expected to be selected again as taste variables in the $\boldsymbol{Z}$-block. As can be seen in Figure 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 $\boldsymbol{X}$ than in $\boldsymbol{Z}$. This means that the variation that is “left” in $\boldsymbol{Z}$ is less important, since some of it is already accounted for by $\boldsymbol{X}$.

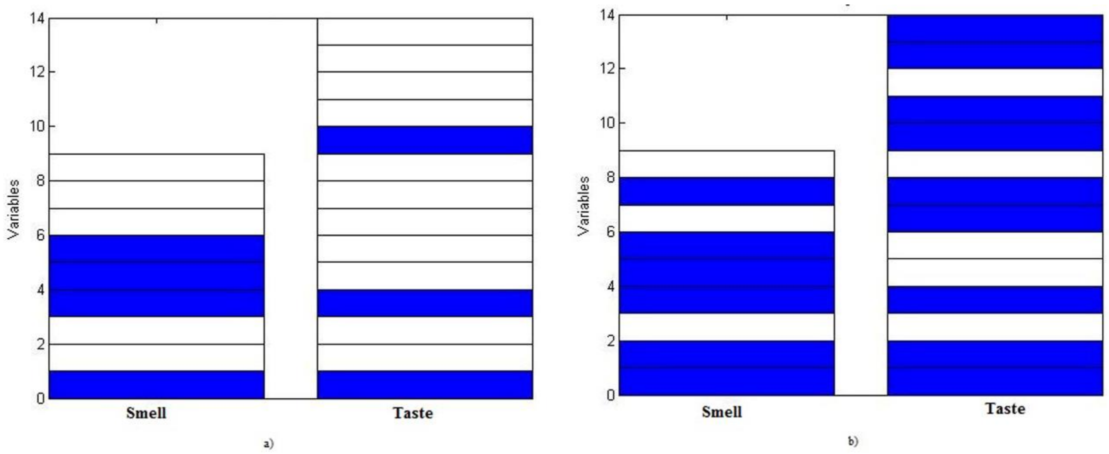

***Figure 1: Selected variables in X- and Z-blocks by variable selection Integrated into SO-PLS models Selected variables are highlighted in blue; (a) variables selected by VIP (b) variables selected by SR.***

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

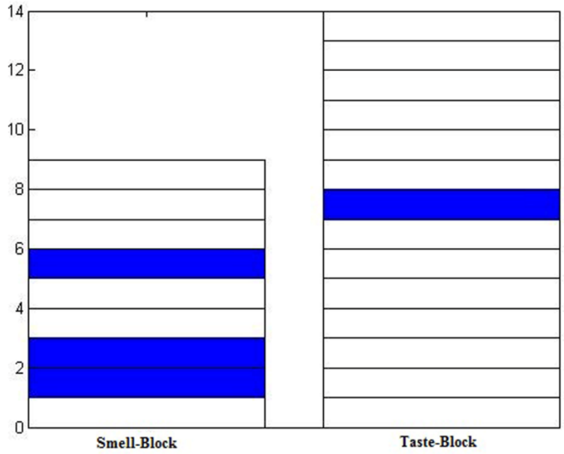

***Figure 2: Selected Variables by the Forward Selection combined with SO-PLS. Selected variables are highlighted in blue.***

## 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 $\boldsymbol{Z}$ alone, and combining $\boldsymbol{X}$ and $\boldsymbol{Z}$ does not improve the prediction ability much. This means that also in this case, the main motivation for doing multi-block analysis is interpretation.

***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 $\mathbf{X}$ | Selected variables $\mathbf{Z}$ | LVs | REMSEP | Explained variance **Y** (%) |
|---|---|---|---|---|---|---|
| No variable selection | Only $\mathbf{X}$ (PLS) | All | None | 3 | 1.61 | 86 |
| | Only $\mathbf{Z}$ (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 |
| | $\text{SR}_{\text{(mean)}}$ | 157/523 | 202/574 | 3 | 2.72 | 62 |
| *Selection on individual block + SO-PLS* | VIP | 182/523 | 136/574 | 3,4 | 1.07 | 96 |
| | SR | 52/523 | 65/574 | 1,7 | 0.79 | 97 |
| | $\text{SR}_{\text{(mean)}}$ | 152/523 | 193/574 | 3,2 | 1.16 | 93 |
| *Selection integrated in SO-PLS* | VIP | 182/523 | 129/574 | 4,5 | 1.24 | 96 |
| | SR | 52/523 | 53/574 | 4,1 | 1.30 | 95 |
| | $\text{SR}_{\text{(mean)}}$ | 152/523 | 102/574 | 3,2 | 1.09 | 94 |
| | Forw. Sel. | 52/523 | 29/574 | 4,1 | 1.19 | 94 |

In order to perform the forward selection on the spectroscopic data set, the training set (both $\boldsymbol{X}$ and $\boldsymbol{Z}$) 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 described in paragraph 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 $\boldsymbol{X}$-block and one interval from the $\boldsymbol{Z}$-block. This amounts to 52 variables from the $\boldsymbol{X}$-block and 29 from the $\boldsymbol{Z}$-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 $\boldsymbol{X}$ and $\boldsymbol{Z}$, regardless of the variable selection method.

Looking more into the selected variables, VIP and SR select different variables from the two blocks. In Figure 3 one can see which variables were selected by VIP, SR and forward selection when integrated into the SO-PLS model. Figures 3(a) and 3(b) represent spectra in $\mathbf{X}$ and $\mathbf{Z}$, 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 bold face variables are those selected by VIP. The lines at the bottom (offset downwards) show in bold face the variables selected by the forward selection. From the interpretation point of view, VIP is the more interesting. Indeed, looking at the fingerprint region, (Figure 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 cm$^{-1}$). For the $\mathbf{Z}$-block (Figure 3b), VIP is able to select the most relevant bands. In fact, selected variables are those around 1263 cm$^{-1}$, where the symmetric rocking of $=C-H$ takes place. Moreover, it selects variables around 1445 cm$^{-1}$ where the $CH_2$'s scissoring takes place, and variables around 1656 cm$^{-1}$ where there are the $C=C$ $_{\text{(cis)}}$ stretching and amide I absorptions.

When variable selection is done beforehand on the individual blocks (Procedures 3 and 4 in the list in Paragraph 3.1), VIP selects only seven variables more that those selected following the procedure 5 in Paragraph 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 $\boldsymbol{X}$-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 B.3, SR selects less ‘relevant but not selective’ variables than VIP. Consequently, the behaviors observed is 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 (Figure 3(a), bottom line), it selects variables related to the skeletal $C-O$ stretching (around 925 cm$^{-1}$). Looking at the rest of the Raman spectra (Figure 3(b), bottom line) it picks the $CH_2$’twisting and the $=CH$ ’ symmetric rocking (variables between 1200 and 1356 cm$^{-1}$ ).

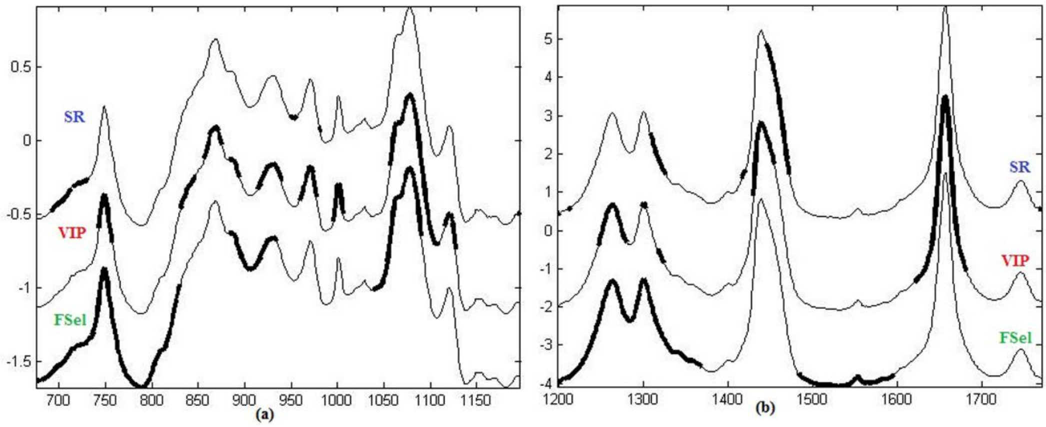

***Figure 3: Selected variables when VIP, SR and Forward Selection are implemented in SOPLS (procedures 5-7). (a) Lines represent the average spectra in $X$. The upmost lines are the average spectra (offset to make them more visible) where the selected variables by SR are bolded. The middle lines are 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 $Z$.***

## 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 Figure 4.

*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.

*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 $\boldsymbol{X}$- and $\boldsymbol{Z}_{\boldsymbol{orth}}$-PLS-scores and loadings [13,14]. After $\boldsymbol{Y}$ is fitted to $\boldsymbol{X}_{\boldsymbol{Red}}$, $\boldsymbol{Z}$ is orthogonalized with respect to the scores of this regression. Consequently, $\boldsymbol{Z}_{\boldsymbol{orth}}$ only contains information not present in $\boldsymbol{X}_{\boldsymbol{Red}}$. Interpreting the $\boldsymbol{Z}_{\boldsymbol{orth}}$-PLS-scores means interpreting the $\boldsymbol{Z}$-block without the redundant information already present in $\boldsymbol{X}_{\boldsymbol{Red}}$. Since the $\boldsymbol{Z}_{\boldsymbol{orth}}$-block is less complex than the $\boldsymbol{Z}$-block, it is easier to interpret.

*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.

*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 $\boldsymbol{X}$; one of these have never been selected from the other methods. Concerning the $\boldsymbol{Z}$-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.

*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 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.

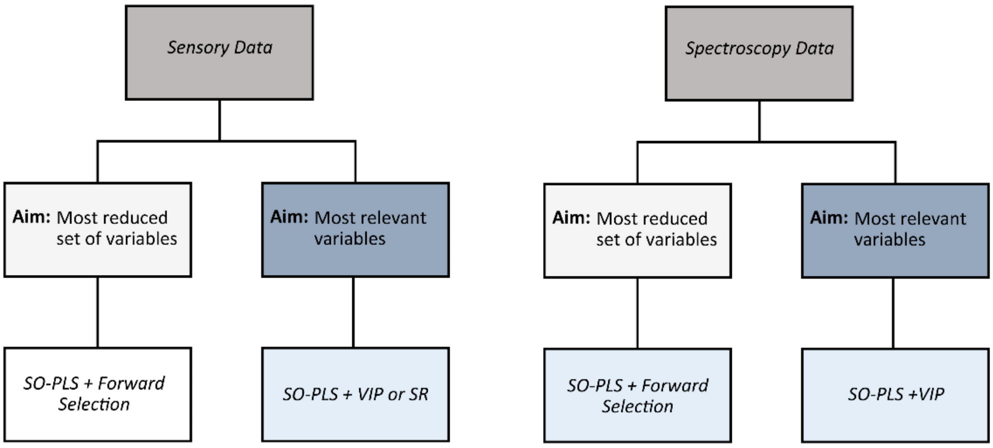
Sensory Data
Spectroscopy Data
Aim: Most reduced set of variables
Aim: Most relevant variables
Aim: Most reduced set of variables
Aim: Most relevant variables
SO-PLS + Forward Selection
SO-PLS + VIP or SR
SO-PLS + Forward Selection
SO-PLS +VIP

***Figure 4: Suggested variable selection approaches for Sensory data and Spectroscopy data***

## 7. Acknowledgements

We would like to thank FFL 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 Paragraph 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 ($\boldsymbol{X}$ and $\boldsymbol{Y}$) and a test set ($\boldsymbol{X_t}$ and $\boldsymbol{Y_t}$). The number of samples ($N$) in the training set is defined according to an experimental design. The $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices have dimensions $N \times 400$ and $N \times 1$, respectively. The $\boldsymbol{X}$ matrix is generated as $\boldsymbol{T_x P_x^T}$. The $\boldsymbol{X}$-scores $\boldsymbol{T_x}$ are simulated from the normal distribution *N*(0,1). The construction of $\boldsymbol{P_x}$ is explained in detail below. The $\boldsymbol{X}$–block is designed as a five-components ($K$) system, hence the dimensionality of $\boldsymbol{T_x}$ will be $N \times 5$. For the scope of this work, it is natural that only some of the components will later contribute to $\boldsymbol{Y}$; 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' ($Ksel$) and the others will be called 'irrelevant components' ($Kirr$). Therefore, the $\boldsymbol{T_x}$ is built as the concatenation of $\boldsymbol{T_{Xsel}}$ and $\boldsymbol{T_{Xirr}}$ scores, where $\boldsymbol{T_{Xsel}}$ represents the 'selective scores' based on the 'selective components', and $\boldsymbol{T_{Xirr}}$ represents the 'irrelevant' ones. These two matrices will have dimensions $(N \times Ksel)$ and $(N \times Kirr)$, respectively. Consequently, the $\boldsymbol{T_x}$-matrix is built as:

$$\boldsymbol{T_x} = [\boldsymbol{T_{Xsel}}\, \boldsymbol{T_{Xirr}}] \qquad (A.1)$$

Then, the coefficient vector $\boldsymbol{b}$ $(Ksel \times 1)$ is generated as a matrix containing random values drawn from the uniform distribution in the open interval (0.05, 1.05).
The response $\boldsymbol{Y}$ is built as:

$$\boldsymbol{Y} = \boldsymbol{T_{Xsel}} * \boldsymbol{b} \qquad (A.2)$$

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 $\boldsymbol{X}$-loadings ($\boldsymbol{P_x}$). 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 Figure A.1. All the different types of variables that constitute the loadings are variables generated using the 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 vary 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 $Msel, Mirr, Mrns$ and $Me$. The 'relevant but not selective'-loadings matrix of dimension $(K \times Mrns)$ is denoted by $\boldsymbol{P_{rns}}$, the 'selective'-loadings matrix of dimension $(Ksel \times Msel)$ is denoted by $\boldsymbol{P_{sel}}$, the 'irrelevant'-loadings matrix of dimension $(Kirr \times Mirr)$ is denoted by $\boldsymbol{P_{irr}}$ and the part representing the noise variables by $\boldsymbol{P_{Noise}}$.

The $\boldsymbol{P_{rns}}$ is a block matrix of the form:

$$\boldsymbol{P_{rns}} = \begin{bmatrix} \boldsymbol{P_{rns}^{sel}} \\ \boldsymbol{P_{rns}^{irr}} \end{bmatrix} \qquad \textit{(A.3)}$$

where $\boldsymbol{P_{rns}^{sel}}$ $(Ksel \times Mrns)$ and $\boldsymbol{P_{rns}^{irr}}$ $(Kirr \times Mrns)$ are matrices of random numbers normally generated. Performing the TP-product to create the $\boldsymbol{X}$-block, the sub-matrix $\boldsymbol{P_{rns}^{sel}}$ will be multiplied by $\boldsymbol{T_{Xsel}}$, while $\boldsymbol{P_{rns}^{irr}}$ is the part multiplied by $\boldsymbol{T_{Xirr}}$. This creates an overlapping between the 'selective' and the 'irrelevant' information. **The** $\boldsymbol{P_{irr}}$ and $\boldsymbol{P_{sel}}$ are in the form:

$$\boldsymbol{P_{irr}} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{P_{irr}^{uniq}} \end{bmatrix} \text{ and } \boldsymbol{P_{sel}} = \begin{bmatrix} \boldsymbol{P_{sel}^{uniq}} \\ \boldsymbol{0} \end{bmatrix} \qquad \textit{(A.5)}$$

where $\boldsymbol{P_{irr}^{uniq}}$ has dimensions $(Kirr \times Mirr)$ and $\boldsymbol{P_{irr}}$ will be of dimensions: $(K \times Mirr)$. $\boldsymbol{P_{sel}^{uniq}}$ has dimension $(Ksel \times Msel)$ and $\boldsymbol{P_{sel}}$ will be of dimensions $(K \times Msel)$. The $\boldsymbol{P_{Noise}}$ consist of zeros only.

This means that $\boldsymbol{P_x}$ can then be represented as a partitioned matrix of the form:

$$\boldsymbol{P_X^T} = \begin{bmatrix} \boldsymbol{P_{rns}} & \boldsymbol{P_{irr}} & \boldsymbol{P_{sel}} & \boldsymbol{P_{Noise}} \\ (K \times Mrns) & (K \times Mirr) & (K \times Msel) & (K \times Me) \end{bmatrix} \qquad \textit{(A.4)}$$

Figure A.1 gives a graphical illustration of how the loadings $\boldsymbol{P_X}$ are partitioned.

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

$$\boldsymbol{X} = \boldsymbol{T_X P_X^T} \qquad \textit{(A.6)}$$

Noise is added to the $\boldsymbol{X}$- and $\boldsymbol{Y}$-blocks. For $\boldsymbol{Y}$, the noise corresponds to a certain percentage of the standard deviation of $\boldsymbol{Y}$ as reported below in Appendix B. For $\boldsymbol{X}$, the standard deviation for each column of $\boldsymbol{X}$ 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 $\boldsymbol{X}$-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 $(Nt)$ is higher. The dimensionality of $\boldsymbol{X_t}$ and $\boldsymbol{Y_t}$ is fixed; these are $1000 \times 400$ and $1000 \times 1$, respectively. The $\boldsymbol{X}$-scores for the test set $\boldsymbol{T_{Xtest}}$, are generated as before and have dimensions $(Nt \times K)$.

The distinction among the variables that has been defined for the training set also applies to the test sets. $\boldsymbol{Y_t}$ is calculated by the *selective scores* for the test set, $\boldsymbol{T_{Xselt}}$ :

$$\boldsymbol{Y_t} = \boldsymbol{T_{Xselt}} * \boldsymbol{b} \qquad \textit{(A.7)}$$

Since the loadings are the same as in the training set, $\boldsymbol{X_t}$ is calculated as:

$$\boldsymbol{X_t} = \boldsymbol{T}_{Xtest}\boldsymbol{P}_X^T \qquad \textit{(A.8)}$$

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 $\boldsymbol{X}$ in *Dataset-2* varies following the design described in Appendix B, while the number of columns is fixed to 40.

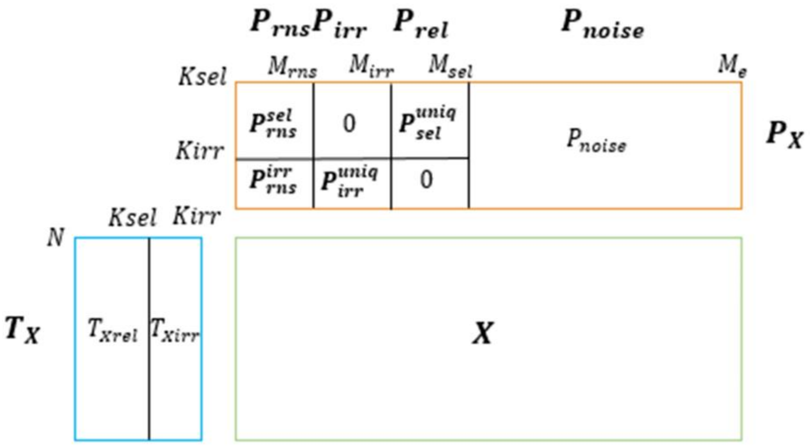

***Figure A.1 : Graphical representation of the simulation of the matrices X, T_X and P_X. The figure shows the partition of $P_X$ in 'relevant but not selective'-variables $P_{rns}$, 'irrelevant'-variables $P_{irr}$, 'selective'-variables $P_{sel}$, and noise-variables P_noise; and their specific dimensions. $P_{rns}$, $P_{irr}$ and $P_{sel}$ are partitioned matrices. $P_{rns}$ is constituted by the concatenation of $A_{sel}$ (Ksc×Mrns) and $A_{irr}$ (Kirr×Mrns). $P_{irr}$ is partitioned in a submatrix of zeros and $B_{irr}$ (Kirr×Mirr). $P_{sel}$ is partitioned in $B_{sel}$ (Ksel×Msel) and a submatrix of zeros. $T_X$-scores matrix is made by the concatenation of $T_{XSel}$ (N×Ksel) and $T_{Xirr}$ (N×Mirr). More details on the submatrices can be found in the text.***

## 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' (Msel)*
5. *Number of 'irrelevant variables' (Mirr)*
6. *Noise added to the* $\boldsymbol{Y}$ *vector*
7. *Noise added to* $\boldsymbol{X}$

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.

**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 (Paragraph 5.1). Both cut-off values have been used and compared for the flavored waters data set (Paragraph 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 $\boldsymbol{X}$ is explained by each variable and, at the same time, of the $\boldsymbol{X}$'s correlation with $\boldsymbol{Y}$. 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 $\boldsymbol{Y}$ on $\boldsymbol{X}$ [22]. Those are then compared to the regression coefficients of a second regression, in which $\boldsymbol{Y}$ is fitted to an $\boldsymbol{XR}$ 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 Y vector, Noise added to 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 $\boldsymbol{Y}$ (%)* | *Noise of $\boldsymbol{X}$ (%)* |
|---|---|---|---|---|---|---|
| ***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 ($3^6$*8) experiments are simulated for *Dataset-1* and *512* ($2^6$*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 $\boldsymbol{Y}$
- Relative percentage of 'selective' *variables* selected (*Rsel*)
- Relative percentage of 'irrelevant*' variables* selected (*Rirr*)
- Relative percentage of the 'relevant but not selective' variables selected (*Rrns*)
- Relative percentage of noise-variables selected (*Rnoise*)
- Relative percentage of total variables selected (*Rtot*)

#### ***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 $\boldsymbol{X}$.

Averaged RMSEPs, *Rsel, Rirr, Rrns* and *Rtot* 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.

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 $\boldsymbol{Y}$-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 $\boldsymbol{Y}$ is at the lower level (15% of the standard deviation of $\boldsymbol{Y}$), 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 modelled is actually captured by the models. Similarly, when the noise in $\boldsymbol{Y}$ 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 $\mathbf{Y}$. 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 $\mathbf{X}$ and in $\mathbf{Y}$ 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 $\mathbf{Y}$. 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.

***Table B.2: Dataset-1: Means (over all the experiments) of RMSEP, Rrns, Rirr, Rsel, and Rtot for each variable selection method.***

| | RMSEP | *Rrns* | *Rirr* | *Rsel* | Rtot |
|---|---|---|---|---|---|
| **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 |

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 $\boldsymbol{X}$* 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 interferents, 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.

### *Selection of the most appropriate variable selection method*

### *Desirability index*

The desirability index (*di*) proposed here is based on the relative percentage of 'selective' *variables (Rsel),* relative percentage of 'irrelevant' variables *($Rirr$), r*elative 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*.

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.

***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.***

| Method | VIP | SR | Jackknifing | sMC | tPLS | UVE |
|---|---|---|---|---|---|---|
| *di* | **0.77** | **0.84** | 0.60 | 0.55 | 0.72 | 0.65 |

## 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 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.

## 8. 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, Revista Colombiana de Estadistica 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, Chemometr. Intell. Lab. Syst. 104 (2010) 140–153.

[4] S. Wold, S. Hellberg, T. Lundstedt, M. Sjostrom and 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. Psychometr.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, Chemometr. 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. Pref. 24 (2012) 8–16.

[8] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemometr. 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 Bioinformatics 15 (2014) 239.

[10]C. M. Andersen and R. Bro, Variable selection in regression—a tutorial, J. Chemometr. 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, Chemometrics and Intelligent Laboratory Systems 118 (2012) 62-69.

[12]J.A. Westerius, T. Kourti, J.F. MacGregor, Analysis of hierarchical PCA and PLS models, J. Chemometr. 12 (1998) 301–321

[13]T. Næs, O. Tomic, B.-H. Mevik, H. Martens, Path modelling by sequential PLS regression, J. Chemometr. 25 (2011) 28–40.

[14]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.

[15]S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures. 3D QSAR in drug design 1 (1993) 523–550.

[16]I.G. Chong, C.H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemometrics and Intelligent Laboratory Systems 78 (2005) 103-112.

[17]S. Favilla, C. Durante, M. Li Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemometrics and Intelligent Laboratory Systems 129 (2013) 76-86.

[18]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. Chemometr. 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. Chemometr. 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), Chemometr. 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, Chemometr. Intell. Lab. Syst. 122 (2013) 103–111.

[24]N. R. Draper, H. Smith, Applied Regression Analysis. Hoboken, NJ: Wiley-Interscience, (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. Pref. 11 (2000) 5-16.

[27]U. Indahl, T. Næs, Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling, J. Chemometr. 12 (1998) 261–278.

[28]L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and 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.