Elsevier Editorial System(tm) for
Chemometrics and Intelligent Laboratory Systems
Manuscript Draft

Manuscript Number: CHEMOLAB-D-16-00301R1

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

Article Type: VSI:multiway/multiset 2016

Section/Category: General Section

Keywords: Multi-block; SO-PLS; multi-way; MB-PLS; classification

Corresponding Author: Dr. Alessandra Biancolillo, M.Sc.

Corresponding Author's Institution: Nofima AS

First Author: Alessandra Biancolillo, M.Sc.

Order of Authors: Alessandra Biancolillo, M.Sc.; Tormod Næs, Prof.;
Rasmus Bro, Professor; Ingrid Måge, Dr.

Reviewer #1: The idea behind the proposed methodology appears to be sound, i.e., to combine the potentiality of multi-block and multi-way analyses, in order to produce a better model for both regression and classification.
The main problem with this manuscript is that the performance improvement announced in the introduction is really not seen in the experimental data sets. On the contrary, previous methods appear to be better for regression and classification than the newly proposed one. The authors recognize this, and provide an explanation based on the increased flexibility of unfolded methods to handle non-linearities. This may be or may not be the case, but the important fact here is that the new method does not perform better than the old ones, as promised.
In any case, non-linearities should not be judged by visual inspection, but on statistical tests. See, for example, S.V.C. de Souza, R.G. Junqueira, A procedure to assess linearity by ordinary least squares method, Anal. Chim. Acta 552 (2005) 25-35, and V. Centner, O.E. de Noord, D.L. Massart, Detection of nonlinearity in multivariate calibration, Anal. Chim. Acta 376 (1998) 153-168.

Another problem is the small number of experimental samples, which precludes reaching a definite and reliable conclusion. For the experimental data sets, the authors only report the cross-validation RMSECV results. It is desirable to have more samples in order to test the results on independent specimens. Otherwise, the conclusions may not be reliable. Perhaps if more samples are analyzed, better conclusions could be reached regarding the alleged superiority of the new method over competitors.

As the reviewer 1 properly affirms, SO-N-PLS does not always perform better than the other presented methods in terms of prediction. In general, results are comparable; in few cases, another method gives slight better results. The main advantage of the novel method is that it allows handling (at the same time) multi-way arrays of data avoiding unfolding. It is therefore expected to give models easier to interpret than models based on unfolded data because of a more parsimonious model. Parsimonious models are always to be preferred because of the parsimony. Consequently, its relevance is on the interpretation of the models even if the accuracy of the predictions is not improved. The introduction has been slightly modified in order to make clearer that the novel method can be used for regression and classification but that its potentialities are more related to the interpretation of the models rather than to the prediction of a response. The analysis of two additional real data sets has been added to the paper in order to discuss more extensively the potential of the novel method.

Following the reviewer suggestion, the discussion of the linearity between the Z-components from SO-PLS and SO-N-PLS and the response (Section 4.2 in the manuscript) has been extended and further investigated. (Non)-linearities have been tested by the run test and the Durbin Watson test. The Durbin Watson test indicates that the second Z-component from SO-PLS is the only one having a linear relationship with the concentration of propanol. From the run test it appears that the second and (to a lesser extent) the first Z-components from SO-PLS have a linear relationship with the response. The Z-components from the SO-N-PLS model never show a linear relation with the response. These results have been included in the manuscript.

Minor:
1) Do not use unexplained acronyms, especially in the title. In the Introduction, EEM, etc. need to be detailed the first time they are introduced.

All the acronyms have been removed or explained.

2) Even when Figure 1 is clear on the advantage of using SO-NPLS over other procedures, at least in the simulations and for some noise levels, I think you need to compare the results using some statistical test with a certain significance, and not leave the comparison to visual inspection. You may use the randomized test of van der Voet (Chemom. Intell. Lab. Syst. 25 (1994) 313) to compare RMSE values. ANOVA analysis of methods, samples and noise would tell you the relative importance with regard to each other, which is OK, but will not tell if the gain in RMSE with one particular method over a competitor is statistically significant.

Following the suggestion of the reviewer, the qualitative considerations already reported in the manuscript have been confirmed using the randomization approach of Van der Voet. The following sentence has been added to the manuscript: “These qualitative considerations based on visual inspection of the figures have also been confirmed by statistical testing using the randomization approach suggested in [30]. The test showed that the differences between the results obtained with SO-N-PLS and either SO-PLS or MB-PLS, for the cases with high noise and a small number of samples, were statistically significant ($p<0.05$). On the other hand, no significant difference was evidenced between the results of SO-PLS and MB-PLS.”

Reviewer #2: The article entitled "Extension of SO-PLS to multi-way arrays: SO-N-PLS" provides a good description of a novel method for regression/classification in the presence of multiple blocks of data, being either some or all of them multi-way structured. Overall, the proposed approach seems to be sound and the comparison carried out by the authors clearly proves the pros of SO-N-PLS over MB-PLS and SO-PLS in specific circumstances. I consider the manuscript acceptable for publication after minor revisions:

- I would slightly rewrite the introduction of the paper. First, I would stress more the issue of the parameter estimation. This is a really interesting problem and, according to me, needs further comments. Please also refer to the article "Bilinear modeling of batch processes. Part III: parameter stability", in which an interesting overview of such an aspect can be found. Then, I would better highlight the advantages resulting from fusing multiple blocks of data. Both these points are fundamental to justify the need, the importance and the use of SO-N-PLS and deserve much more attention to be properly elucidated;

Both the suggested aspects are now mentioned in the introduction. Firstly, the relevance of applying a data-fusion technique handling multi-block data sets is mentioned. Then, SO-N-PLS is presented as “a suitable solution for issues related to parameter stability, e.g. in process monitoring” and the reader is addressed to the suggested paper.

- Sometimes, the text is a bit hard to follow (see e.g. Section 2.3 and 2.4). I think the authors should improve the general readability of the article;

In the entire manuscript, some parts have been re-written in order to increase clarity.

- The mathematical notation is not coherent along the manuscript. I would suggest to always use lower-case italic characters for scalars, lower-case bold characters for vectors and upper-case bold characters for matrices;

The manuscript has been fixed according to the reviewer suggestions.

- I think the authors should also justify the choice/need of applying LDA in the PLS subspace. I guess that could be of interest for potential readers.

The reasons for the choice have been briefly summarized in the manuscript and the reader is addressed to a reference for more details.

Highlights (for review)

HIGHLIGHTS

- Extension of SO-PLS to multi-way arrays. SO-N-PLS handles multi-way arrays: unfolding is not required.

- SO-N-PLS filters out the noise better than SO-PLS and MB-PLS. Simulation studies show that SO-N-PLS performs better than the unfolded methods (SO-PLS and MB-PLS) when the sample size is small and the data is noisy.

- SO-N-PLS gives rise 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.

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

Alessandra Biancolillo [a,b,*], Tormod Næs[a,b], Rasmus Bro[b], Ingrid Måge[a]

[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*

*Corresponding author:
Tel: +47 64 97 01 15
e-mail: 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, were 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. 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 show that SO-N-PLS predicts 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 real data examples, the differences in prediction are small but the multi-way method allows easier interpretation.

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*. This type of data structure 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 -Nuclear Magnetic Resonance (NMR), fluorescence Excitation-Emission Matrix, (EEM)-, chromatography -Gas Chromatography (GC), Liquid Chromatography coupled with Mass Spectrometry (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 r*ow-wise unfolding, column-wise unfolding* and *tube-wise unfolding*. Applying the first approach, a three-way array $\underline{\boldsymbol{X}}$ of dimensions $N \times J \times K$ can be unfolded to a matrix $\boldsymbol{X}$ of dimensions $N \times (JK)$. In the *column-wise* approach, a three-way array $\underline{\boldsymbol{X}}$ of dimensions $N \times J \times K$ is unfolded to a matrix $\boldsymbol{X}$ of dimensions$(NK) \times J$. Finally, in the *tube-wise* approach, the $\boldsymbol{X}$ 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.

The development of new technologies and instrumentations is also making it common to have *multi-block data sets*. For example, the data blocks could represent background information on the samples, various instrumental measurement techniques, different times of measurement and multiple quality attributes. In these situations, it is more efficient to analyse the blocks jointly in order to account for all the actual information on the system at study [8-9].

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) [10-16].

This is a new and emerging field, and many research challenges remain unsolved. Previous work [13,17] 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 without unfolding the multi-way arrays. This characteristic provides a number of advantages. First of all, SO-N-PLS is expected to lead to models easier to interpret than unfolded analysis and may have a better prediction accuracy. Additionally, it could represent a suitable solution for issues related to parameter stability, e.g. in process monitoring (a proper description of the problem falls outside the scope of this paper, but it can be found in [18]).

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 give improved interpretation compared to unfold analysis.
- When three-way data arrays present a clear trilinear structure, 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.

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

--*Figure 1 approx. here*-------------------------------------

2 Material and methods

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

Sequential and Orthogonalized Partial Least Squares (SO-PLS) [13] 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:

$$\mathbf{Y} = \mathbf{Xg} + \mathbf{Zh} + \mathbf{E} \qquad (1)$$

where $\boldsymbol{X}_{(N\times J)}$ and $\boldsymbol{Z}_{(N\times M)}$ are the two predictors blocks, $\boldsymbol{Y}_{(N\times R)}$ 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 [13].

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 $\mathbf{Z}_{\text{orth}}$):

$$\mathbf{Z_{orth}} = \mathbf{Z} - \mathbf{T_X}(\mathbf{T_X}^{\mathbf{T}}\mathbf{T_X})^{-1}\mathbf{T_X}^{\mathbf{T}}\mathbf{Z} \qquad (2)$$

3. Residuals from the first PLS are fitted to $\mathbf{Z}_{\text{orth}}$, giving PLS scores $\mathbf{T_{Zorth}}$.

4. The full predictive model is computed as the ordinary least squares fit of $\mathbf{Y}$ to $\mathbf{T}_{\mathbf{X}}$ and $\mathbf{T}_{\mathbf{Zorth}}$.
   Since the first set of scores are linear functions of $\mathbf{X}$ and the second set of scores are linear functions of $\mathbf{Z}_{\mathrm{orth}}$ 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 [1]. 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 $\mathbf{T}_{\mathbf{Zorth}}$ obtained in step 3 and the loadings obtained by regressing $\mathbf{Z}$ onto $\mathbf{T}_{\mathbf{Zorth}}$.
SO-PLS can be combined with Linear Discriminant Analysis (LDA) [19] in order to create classification models [17]. A discriminant classifier has been preferred because it ensures that each sample is assigned to only one class. Among the discriminant classification methods, LDA has been chosen because it fits well with the SO-PLS philosophy and it allows representation of the classification results by means of canonical variates (more details in [17]). The method is then called SO-PLS-LDA, and the only difference is that LDA is applied to the concatenated scores $[\mathbf{T}_{\mathbf{X}}\ \mathbf{T}_{\mathbf{Zorth}}]$ 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 [10-11]. 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 [10-11]. 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 [20]; we are going to refer to this method as MB-PLS-LDA.

2.3 N-PLS

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:

$$x_{njk} = \sum_{f=1}^{F} t_{nf} w_{jf}^{J} w_{kf}^{K} + e_{njk} \qquad (3)$$

where $\boldsymbol{t}$ are the scores and $\mathbf{w}^{\mathrm{J}}$ and $\mathbf{w}^{\mathrm{K}}$ are the weights of the second and third mode, respectively. The model corresponds to the so-called Martens PLS algorithm [21], in which 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 the two sets of loading weights $\mathbf{w}$ provide scores $\boldsymbol{t}$ 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. 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 $\mathbf{T_{Zorth}}$) 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* x *M* x *I* is first unfolded row-wise to $\mathbf{Z}_{\mathrm{un}}$, a matrix of dimensions $N \times MI$. Then, $\mathbf{Z}_{\mathrm{orth}}$ is obtained replacing $\mathbf{Z}$ with $\mathbf{Z}_{\mathrm{un}}$ in Eq.2 before $\mathbf{Z}_{\mathrm{orth}}$ 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:

$$\mathbf{Y} = \mathbf{X}_{\mathrm{un}}\boldsymbol{\gamma} + \mathbf{Z}_{\mathrm{un}}\boldsymbol{\nu} + \mathbf{E} \qquad (4)$$

Where $\mathbf{X}_{\mathrm{un}}$, is the unfolded $\underline{\mathbf{X}}$, a matrix of dimensions $N \times JK$. $\boldsymbol{\gamma}_{(JKxR)}$ and $\boldsymbol{\nu}_{(MIxR)}$ are the regression coefficients. Note that *N*-PLS involves two sets of weights, $\mathbf{w}^{\mathrm{J}}$ and $\mathbf{w}^{\mathrm{K}}$. 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 $\boldsymbol{\nu}$, 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 [22] (procedure called *Method 2*).

Firstly, weights $\boldsymbol{W}$ are calculated as the inner product of the weights $\mathbf{w}^{\mathrm{J}}$ and $\mathbf{w}^{\mathrm{K}}$. Then, the loading-weights $\mathbf{R}$ are obtained as:

$$\boldsymbol{R} = \boldsymbol{W}/\boldsymbol{\delta} \qquad (5)$$

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

Then, $\mathbf{b}_{\mathrm{N-PLS}}$ can be calculated as:

$$\mathbf{b}_{\mathrm{N-PLS}} = \mathbf{R}\boldsymbol{Q}^{\boldsymbol{T}} \qquad (6)$$

where $\boldsymbol{Q}$ are the $\boldsymbol{Y}$-loadings.

SO-N-PLS can be used for classification problems by applying LDA on the concatenated $\mathbf{T}_{\mathrm{X}}$ and $\mathbf{T}_{\mathrm{Zorth}}$, 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 [13]. 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 [17].

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 [13,17]. 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 $\boldsymbol{Z}$ 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 $\mathbf{w}^{\mathrm{J}}$ and $\mathbf{w}^{\mathrm{K}}$ individually. Another alternative is to plot the outer product of $(\boldsymbol{w}^{J}\boldsymbol{w}^{K})^{T}$ as a landscape.

For MB-PLS and SO-PLS, $\boldsymbol{X}$-loading weights for each component will be of size $\boldsymbol{JK} \times \mathbf{1}$. They can be plotted as they are, or folded back to an $\boldsymbol{J} \times \boldsymbol{K}$ matrix and plotted as a landscape similarly to the outer product $\mathbf{w}^{\mathrm{J}}$ and $\mathbf{w}^{\mathrm{K}}$ for SO-N-PLS.

Interpretation of the $\mathbf{Z}$-block is slightly different than for the $\mathbf{X}$-block since $\mathbf{Z}_{\mathrm{orth}}$ 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 $\mathbf{P}_{\mathrm{z}}$ as projections of $\mathbf{Z}$ itself on $\mathbf{T}_{\mathrm{Zorth}}$ [17]:

$$\mathbf{P}_{\mathrm{z}} = (\mathbf{T}_{\mathbf{Zorth}}{}^{\mathbf{T}}\mathbf{T}_{\mathbf{Zorth}})^{-1}\mathbf{T}_{\mathbf{Zorth}}{}^{\mathbf{T}}\mathbf{Z} \qquad (7)$$

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 $\mathbf{T}_{\mathbf{Zorth}}$:

$$\mathbf{W}_{\mathrm{z}} = (\mathbf{T}_{\mathbf{Zorth}}{}^{\mathbf{T}}\mathbf{T}_{\mathbf{Zorth}})^{-1}\mathbf{T}_{\mathbf{Zorth}}{}^{\mathbf{T}}\mathbf{Z}_{\mathrm{un}} \qquad (8)$$

By Eq. 8 we obtain unfolded $\mathbf{W}_\mathrm{z}$. These can then be reshaped and plotted in the same ways as for $\mathbf{W}_\mathrm{x}$.

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 ($\boldsymbol{\gamma}$ 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.

3. Data sets

3.1 Simulated Data

Data sets consisting of two three-way predictor blocks ($\underline{\mathbf{X}}$ and $\underline{\mathbf{Z}}$) 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 6×4=24 different factor combinations. Noise was added to each variable of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Z}}$. 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 × 4 × 100 = 2400 different data sets. For each data set, an independent test set ($\underline{\mathbf{X}}_\mathbf{t}$, $\underline{\mathbf{Z}}_\mathbf{t}$ and $\mathbf{y}_\mathbf{t}$) of 600 samples was constructed for validation purposes.

The three-way $\underline{\mathbf{X}}$, $\underline{\mathbf{Z}}$, $\underline{\mathbf{X}}_\mathrm{t}$ and $\underline{\mathbf{Z}}_\mathrm{t}$ predictor blocks were simulated to mimic fluorescence spectra, and were created in the following way:

$\underline{\mathbf{X}}(\, N_i \times 201 \times 61)$ is generated as the outer product of $\mathbf{T}_\mathbf{X}$, $\mathbf{B}_\mathbf{X}$ and $\mathbf{C}_\mathbf{X}$ while $\underline{\mathbf{Z}}(\, N_i \times 201 \times 61)$ as the outer product of $\mathbf{T}_\mathbf{Z}$, $\mathbf{B}_\mathbf{Z}$ and $\mathbf{C}_\mathbf{Z}$. Scores $\mathbf{T}_\mathbf{X}$ and $\mathbf{T}_\mathbf{Z}$ are both $(N_i \times 2)$ matrices of normally distributed random numbers. $\mathbf{B}_\mathbf{X}$ and $\mathbf{B}_\mathbf{Z}$ (both 201×2), and $\mathbf{C}_\mathbf{X}$ and $\mathbf{C}_\mathbf{Z}$ (both 61×2) are loadings extracted from real fluorescence spectra of mixtures of aminoacids (data set described in [23]). Consequently, the loading vectors are not orthogonal. Correlations between components within each loading are -0.21, -0.48, 0.93 and -0.15,for $\mathbf{B}_\mathbf{X}$, $\mathbf{B}_\mathbf{Z}$, $\mathbf{C}_\mathbf{X}$ and $\mathbf{C}_\mathbf{Z}$, respectively.

The response vector $\mathbf{y}$ is built as:

$$\mathbf{y} = [\mathbf{T}_\mathbf{X}\ \mathbf{T}_\mathbf{Z}] * \boldsymbol{\beta}$$

Where $\boldsymbol{\beta}$ (4×1) is the coefficient vector generated as a matrix containing random values drawn from the uniform distribution on the interval (0.05, 1.05)**.**

$\mathbf{T}_{\mathbf{X}}$, $\mathbf{T}_{\mathbf{Z}}$ 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 [24]. 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 $\mathbf{y}_{\mathrm{VTV}}$, $\mathbf{y}_{\mathrm{TG}}$, $\mathbf{y}_{\mathrm{Phe}}$, $\mathbf{y}_{\mathrm{Mal}}$ and $\mathbf{y}_{\mathrm{Pro}}$ 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 $\boldsymbol{Z}\,(58 \times 9168)$. 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 [25].

3.4 Butter Data set

Butter is a food product widely consumed all over the world. It contains different photosensitizers which absorb in the UV, violet and visible regions of the spectrum. Various samples of industrial butter were analyzed in order to study the oxidation of photosensitizers at different conditions. Samples were packed and exposed to three different light colors (Violet, Green and Red) in different atmospheres (low or high oxygen); two samples were not exposed to any light. The duration of the exposure varied from six to forty-eight hours. Of these, twenty-one samples were analysed by fluorescence spectroscopy (EEM). The emission side was scanned from 580 to 720 nm, and excitation was scanned from 350 to 452 nm. Single emission spectra (405-563 nm) were also measured on a different instrument. This instrument has a higher signal-to-noise ratio and might therefore contain more detailed information than the EEM landscapes. Finally, the same samples were judged by eleven panelists (more details can be found in [26]). Here, the two fluorescence data blocks will be used to predict the sensoric attribute *acidic odour*. The EEM array of dimensions 21×274×35 will be used as $\underline{\mathbf{X}}$, the emission fluorescence block of dimensions 21 × 392 as $\mathbf{Z}$ and the $\mathbf{y}$ vector of dimensions 21 × 1 is the response vector.

3.5 Sugar process data set

Refinement of sugar is a long process divided in different steps. In order to understand the chemistry involved in the process, 268 sugar samples from the last stage of the productive process were measured spectrofluorometrically. Emission spectra were registered from 275 to 560 nm while excitation was registered for seven wavelengths (230, 240, 255, 290, 305, 325, 340 nm), yielding a three-dimensional $\underline{\mathbf{X}}$-data block of dimension 268 × 571 × 7. Additionally, seven auxiliary laboratory measurements $(\boldsymbol{Z}_{(268\times7)})$ were used together with the fluorescence spectra for the prediction of the sugar color $(\boldsymbol{y}_{(268\times1)})$, defined as a unit derived from the absorbance at 420 nm. For more details, please consult the original publication [27].

4. Results and Discussion

The simulated data sets and the *sugar* data set were validated by independent test sets. Sugar samples were divided in training (200 samples) and test set (68 samples) by the Duplex algorithm [28]. The *mixture*, *Lambrusco* and the *butter* 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: P*art 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 χ^2 test (significance level 5%) [29]. *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.

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.

--*Table 1 approx. here*------------------------------------

--*Figure 2 approx. here*------------------------------------

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 approx. here*------------------------------------

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. These qualitative considerations based on visual inspection of the figures have also been confirmed by statistical testing using the randomization approach suggested in [30]. The test showed that the differences between the results obtained with SO-N-PLS and either SO-PLS or MB-PLS, for the cases with high noise and a small number of samples, were statistically significant ($p<0.05$). On the other hand, no significant difference was evidenced between the results of SO-PLS and MB-PLS.

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 approx. here*----------------------------------

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 $\underline{\mathbf{X}}$ and $\underline{\mathbf{Z}}$.

---*Figure 5 approx. here*----------------------------------

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 dataset (the same data set shown in Fig. 5) are shown in Fig. 6.

These coefficients correspond to $\mathbf{g}$ and $\boldsymbol{\gamma}$ 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 approx. here*----------------------------------

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.

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.

--*Table2 approx. here*-----------------------------------

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. From a visual inspection, the second component from SO-PLS appears the only one that has a clear, linear relationship with propanol. In order to investigate this aspect properly, randomness in the residuals have been tested by two different statistical test: the *run test* and the *Durbin Watson test* [31]. According to the run test, only the first and the second $\mathbf{Z}_{orth}$-components from SO-PLS have a linear relation with the propanol concentrations (under the null hypothesis of randomness in residuals, p-values are 0.09 and 0.55, respectively). The Durbin Watson test suggests that only the second $\mathbf{Z}_{orth}$-component from the SO-PLS model has a linear relation with the response (p-value 0.189). These results indicate 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 three-way data structure. The high number of components SO-PLS selected from the $\mathbf{Z}$-block also confirms this hypothesis.

--*Figure 7 approx. here*-----------------------------------

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 $\boldsymbol{X}$ and $\boldsymbol{Z}$, 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 $\mathbf{X}$-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 (Figg. 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 approx. here*-----------------------------------

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.

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 (**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.

---*Table3 approx. here*-----------------------------------

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 $\underline{\boldsymbol{X}}$ (GC-MS) and PLS-LDA on the two-way $\boldsymbol{Z}$ (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 **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 approx. here*-----------------------------------

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 [17].

---*Figure 9 approx. here*-----------------------------------

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 $\underline{\boldsymbol{X}}$, MB-PLS-LDA and SO-PLS-LDA misclassify the same samples, indicating that these are intrinsically hard to distinguish. PLS-LDA on $\boldsymbol{Z}$ misclassifies even more samples (six and three misclassified for Salamino and Sorbara, respectively).

---*Figure 10 approx. here*-----------------------------------

4.4 **Results on the butter data set**

$\overline{\boldsymbol{X}}$ and $\boldsymbol{Z}$ data blocks described in Section 3.4 are used to predict the *acidic odour* of the butter samples. Firstly, the two blocks $\overline{\boldsymbol{X}}$ and $\boldsymbol{Z}$ are used individually for the prediction of the attribute (by N-PLS and PLS, respectively).

Then, SO-N-PLS, MB-PLS and SO-PLS models were built. RMSECVs for all these models are reported in Table 5.

---*Table 5 approx. here*-----------------------------------

Looking at results reported in Table 5, it is evident that the multi-block models show a prediction ability statistically not different from the PLS model on $\boldsymbol{Z}$ (and better than the N-PLS model on $\overline{\boldsymbol{X}}$). From the interpretation point of view, models show a much more different scenario.

Looking at the weights plots reported in Fig. 11 it is quite evident that SO-N-PLS models are the easiest to interpret. Following the identification of peaks suggested in [26], the first component is given by the contribution of *riboflavin* while the second one shows the *chlorophyll* peak. The same compounds would be identified looking at the weights of SO-

PLS (Figg. 11b and 11e), but the peak in Fig. 11e looks less defined than in Fig. 11d. Investigating the MB-PLS $\bar{\boldsymbol{X}}$-weights (Figg. 11c, 11f, 11g and 11h), the interpretation of the peaks would not be straightforward. The first three components seem dominated by the contribution of riboflavin, while the fourth (Fig. 11h) shows a mixed contribution from riboflavin and chlorophyll a.

---Figure 11 *approx. here*------------------------------------

$\boldsymbol{X}$-regression coefficients for the SO-N-PLS, SO-PLS and MB-PLS models are displayed in Fig. 12 a, b and c, respectively. Without going too much in depth in the interpretation of the regression coefficients (which may not be reliable for the reasons discussed in [32]), a visual inspection of Fig. 12 confirms that SO-N-PLS models less noise than the other two methods, as the peaks in Fig. 12a look smoother than those in Fig. 12b and c.

---Figure 12 *approx. here*------------------------------------

4.5 Results on the sugar data set

As described in Section 3.5, the $\bar{\boldsymbol{X}}$ and $\boldsymbol{Z}$ blocks are used to predict the sugar color. Predictions were made by N-PLS and PLS on the individual blocks and by SO-N-PLS, SO-PLS and MB-PLS on both blocks. As mentioned, models were validated on a test set. RMSEPs, explained variances and the number of latent variables used for each model are reported in Table 6.

---Table 6 *approx. here*------------------------------------

From Table 6 we see that multi-block models, in particular SO-PLS, lead to a better prediction ability than models based on the individual data blocks. Even if SO-PLS gives the best predictions, it is not the most informative model from the interpretation point of view. In Fig. 13 the scores plots for the three different multi-block models are reported. Looking at the scores plot from the SO-N-PLS model (Fig. 13 a), it is possible to observe that the first $\boldsymbol{X}$-component is particularly suitable to distinguish the entity of the absorption (and therefore the color) of the different components. It is difficult to conclude the same looking at the scores plots from the models based on the unfolded data (Figg. 13b and c). In fact, in these cases, samples with comparable absorptions are more spread out along the first $\boldsymbol{X}$-component. This is particularly evident looking at the SO-PLS' scores plot, and for samples with $\boldsymbol{y}$ from 20 to 35.

---Figure 13 *approx. here*------------------------------------

5. Discussion of the hypotheses mentioned in the introduction

The first 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 underlining complexity, while MB-PLS and SO-PLS generally need more latent variables (in particular for the $\mathbf{Z}$-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. In all the other sets of data, SO-N-PLS required the same number of components as MB-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 required more latent variables than SO-N-PLS in all cases except one (Lambrusco data set) in which all the methods are selecting the same total complexity. In the further study on this data set with addition of noise to $\mathbf{X}$ and $\mathbf{Z}$, SO-N-PLS confirms its parsimony in the latent variable selection.

Despite the number of latent variable accounted in each SO-N-PLS models, it has been shown (e.g. scores plot in Fig. 13) that these models are more informative (from the interpretation point of view) than the models based on the unfolded data sets.

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.

Additionally, as shown in the weights plot in Figures 8 and 11, SO-N-PLS leads to models whose chemical interpretation would be easier.

The second 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.

From the prediction point of view, 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 $\mathbf{X}$ and $\mathbf{Z}$, SO-N-PLS gives the best predictions. Concerning predictions in the Lambrusco data set, the methods were indistinguishable. Also in the butter data set, results from the three methods are comparable. 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. From the results obtained on the sugar data set, it appears that SO-N-PLS and SO-PLS give comparable predictions outperforming MB-PLS.

5.1 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.

As expected, SO-N-PLS has demonstrated to be the most suitable method for interpretation of multi-way and multi-block data sets. Consequently, it is the suggested approach when interpretation is the main aim of the analysis.

From the prediction point of view, 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 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.

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 (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, Psychometrika, 35 (1970) 283.

[6] R.Bro, multiway calibration. Multilinear PLS, J. Chemomtr. 10 (1996), 47-61.

[7] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometr. 31 (1966) 279–311.

[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] 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.

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

[12] 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.

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

[14] 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.

[15] T. Löfstedt, J. Trygg, OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation, J. Chemometr. 25 (2011) 441–455.

[16] 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).

[17] 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.

[18] J. M. González Martínez, J. Camacho, A. Ferrer, Bilinear modeling of batch processes. Part III: parameter stability, J. Chemometrics 28 (2014) 10–27.

[19] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188.

[20] T. Næs, U. Indahl, A unified description of classical classification methods for multicollinear data. J. Chemometr. 12 (1998) 205–220.

[21] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, New York, 1989.

[22] S. De Jong, Short communication regression coefficients in multilinear PLS, J. Chemometr. 12 (1998) 77-81.

[23] R. Bro, PARAFAC: Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (1997) 149-171.

[24] 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).

[25] 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.

[26] J.P. Wold, R. Bro, A. Veberg, F. Lundby, A.N. Nilsen, J. Moan, Active photosensitizers in butter detected by fluorescence spectroscopy and multivariate curve resolution. J. Agric. Food Chem. 54 (2006), 10197–10204.

[27] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, Chemom. Intell. Lab. Syst. 46 (1999) 133-147.

[28] R.D. Snee, Validation of regression models: methods and examples, Technometrics, 19 (1977) 415–428.

[29] U. Indahl, A twist to partial least squares regression, J. Chemometr.19 (2005) 32–44.

[30] H. van der Voet, Comparing the predictive accuracy of models using a simple randomization test, Chemometr. Intell. Lab. Syst. 25 (1994), 313–323.

[31] N.R. Drapper, H. Smith, Applied Regression Analysis, 2nd ed., Wiley, New York, 1981

[32] K. Kjeldahl, R. Bro, Some common misunderstandings in chemometrics, J. Chemometr. 24 (2010) 558–564.

Figure Captions

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 **X**-block is three-way, while **Z** is a matrix. b) both **X** and **Z**-blocks are three-way arrays. SO-PLS and MB-PLS are always applied on the row-wise unfolded matrices.

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).

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).

Figure 4: Average number of latent variables selected for each level of noise and sample size. 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 $\underline{\boldsymbol{Z}}$**-**block. SO-N-PLS not shown because 2 LVs were always selected for both blocks.

Figure 5: **X**-weights from models on one of the simulated data sets with 60 samples and 50% noise.

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 60 samples and 50% noise.

Figure 7: $\boldsymbol{y}_{Pro}$vs $\boldsymbol{T}_{Zorth}$ from the SO-N-PLS and SO-PLS models: a) and b) $\boldsymbol{y}_{Pro}$ is reported as a function of the first $\boldsymbol{T}_{Zorth}$ from SO-N-PLS and the first $\boldsymbol{T}_{Zorth}$ from SO-PLS, respectively. c) and d) $\boldsymbol{y}_{Pro}$ plotted against the second $\boldsymbol{T}_{Zorth}$ from SO-N-PLS and from SO-PLS, respectively.

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.*

Figure 9: Classification of Lambrusco wines. Predictions in the space on canonical variates using both X (GC-MS) and Z (NMR) blocks. Circled samples are the misclassified ones.

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.

Figure 11 Butter data set: X-weights plots for SO-N-PLS (a and d), SO- PLS (b and e) and MB-PLS (c,f,g and h) models.

Figure 12 Butter data set: Regression coefficients plots from a) SO-N-PLS; b) SO-PLS; c) MB-PLS.

Figure 13 Sugar process data set: Scores plots for the multi-block models built on the sugar data set: a) SO-N-PLS b) SO-PLS (unfolded data) c) MB-PLS (unfolded data). **Legend**: Filled circles for training set samples, empty circles for test set samples. Yellow: $y<20$; Magenta:$20\leq y<25$; Cyan:$25\leq y<30$; Green:$30\leq y<35$; Blue:$35\leq y<40$; Red:$40\leq y \leq 45$.

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	

Table 2: Chemical mixtures data set: RMSECVs and Explained variances for the prediction of the concentrations of compounds in the mixture.

	***N*-PLS (Only X-block)**			***N*-PLS(Only Z-block)**			**SO-*N*-PLS**			**MB-PLS**			**SO-PLS**		
Compound	**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

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 $\boldsymbol{X}$ and $\boldsymbol{Z}$.

	SO-N-PLS			**MB-PLS**		**SO-PLS**		
Compound	**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
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

Table 4: Lambrusco Data set: Classification errors by single-block and multi-block methods.

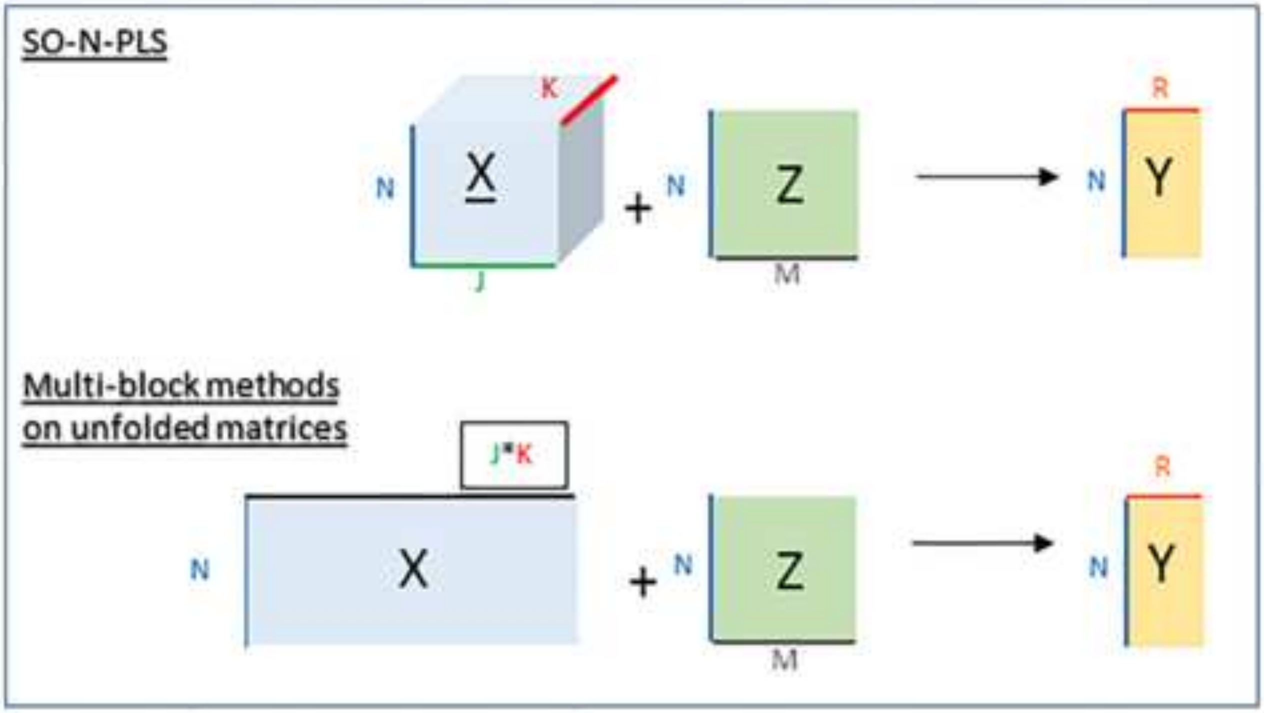
Method	LVs:	Miscl. Grasparossa	Misclassified Salamino	Misclassified Sorbara	Tot.Error (%)
N-PLS-LDA (Only $\overline{\boldsymbol{X}}$)	3	7	4	3	24
PLS-LDA (Only Z)	3	16	13	5	59
SO-N-PLS-LDA	2,4	7	4	3	24
MB-PLS-LDA	6	6	7	2	26
SO-PLS-LDA	2,4	6	5	3	24

Table 5 Butter data set: number of Latent variables (LVs) selected, RMSECVs and Explained variances for the prediction of the acidic odour attribute.

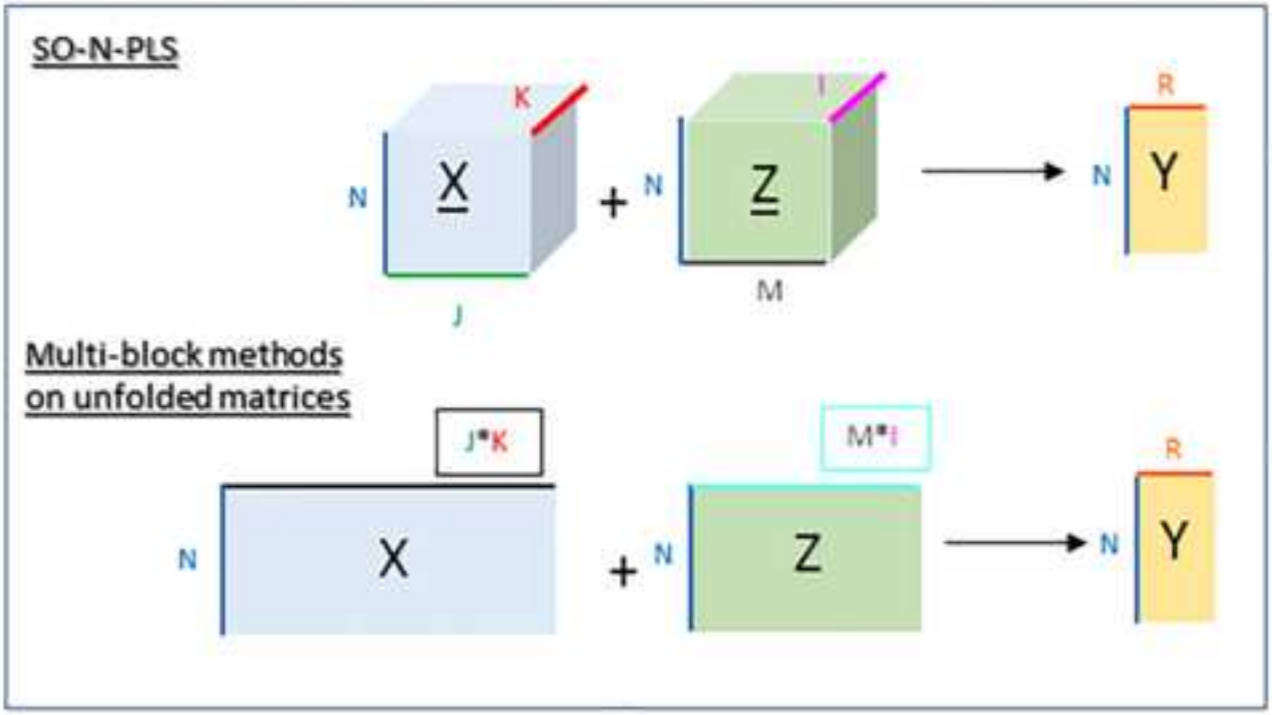
Method	LVs:	RMSECV	Explained Variance (%)
N-PLS (Only $\overline{\boldsymbol{X}}$)	4	0.71	83
PLS (Only $\boldsymbol{Z}$)	5	0.58	88
SO-N-PLS	2,2	0.56	88
MB-PLS	4	0.58	88
SO-PLS	2,4	0.53	88

Table 6: Sugar process data set: number of Latent variables (LVs) selected, RMSECVs and Explained variances for the prediction of the sugar colour.

Method	LVs:	RMSEP	Explained Variance (%)
N-PLS (Only $\bar{\boldsymbol{X}}$)	3	2.27	77
PLS (Only $\boldsymbol{Z}$)	1	4.3	15
SO-N-PLS	2,1	2.23	78
MB-PLS	3	2.36	75
SO-PLS	3,1	2.11	80


SO-N-PLS
K
N
X
J
+
N
Z
M
R
N
Y
Multi-block methods on unfolded matrices
J*K
N
X
+
N
Z
M
R
N
Y

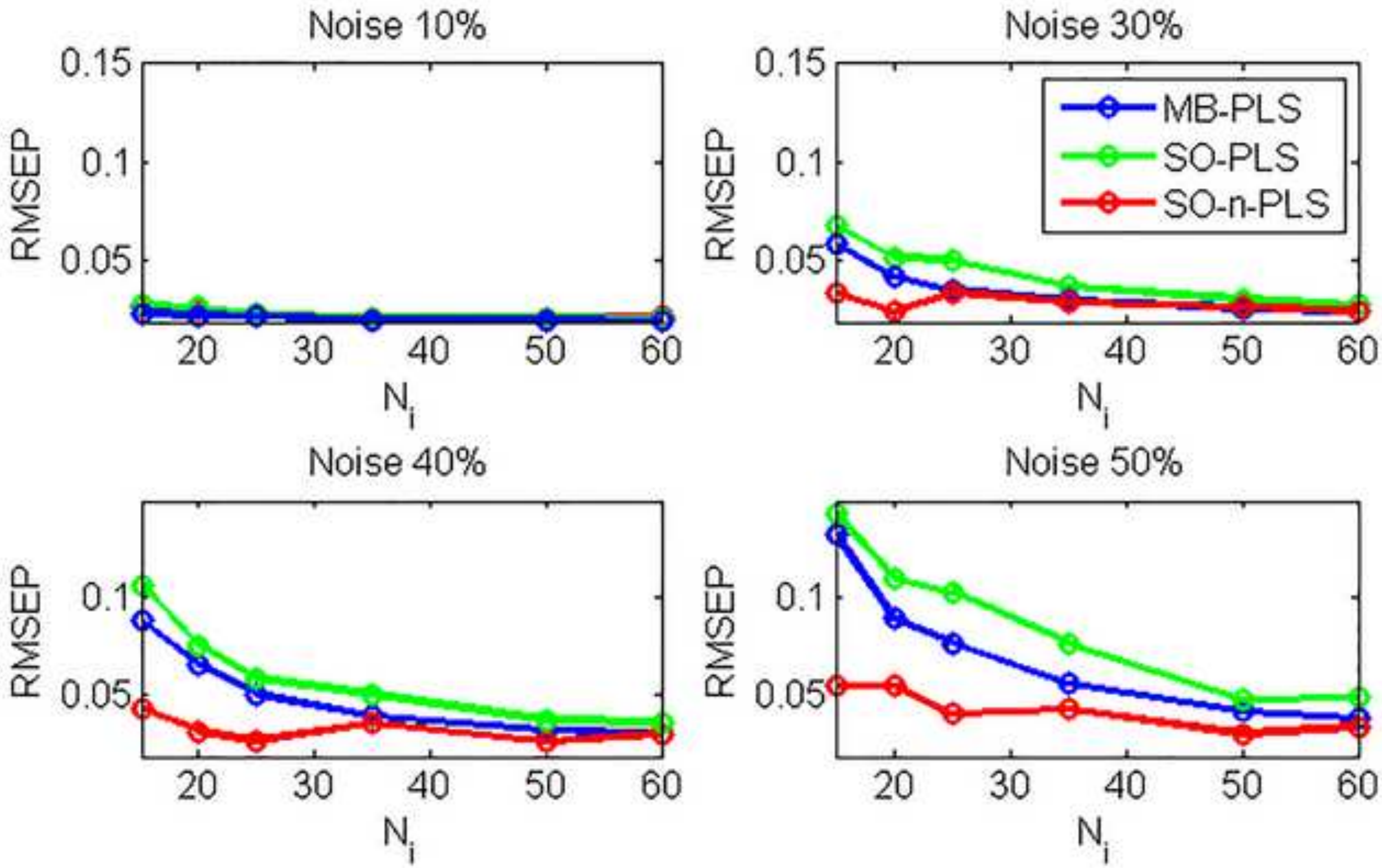
a)


SO-N-PLS
K
N
X
J
+
I
N
Z
M
R
N
Y
Multi-block methods on unfolded matrices
J*K
M*I
N
X
+
N
Z
R
N
Y

b)

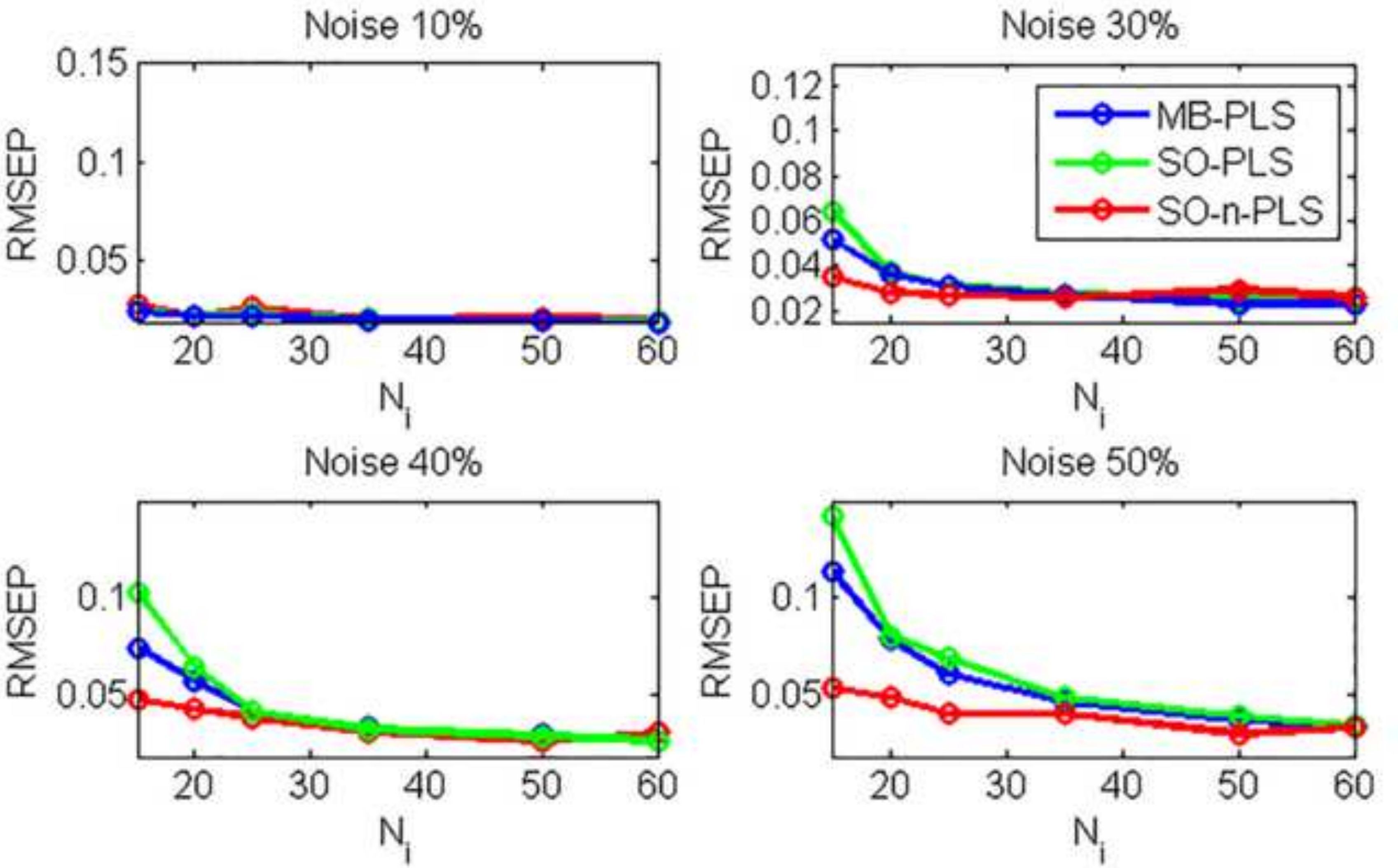
Figure(s)

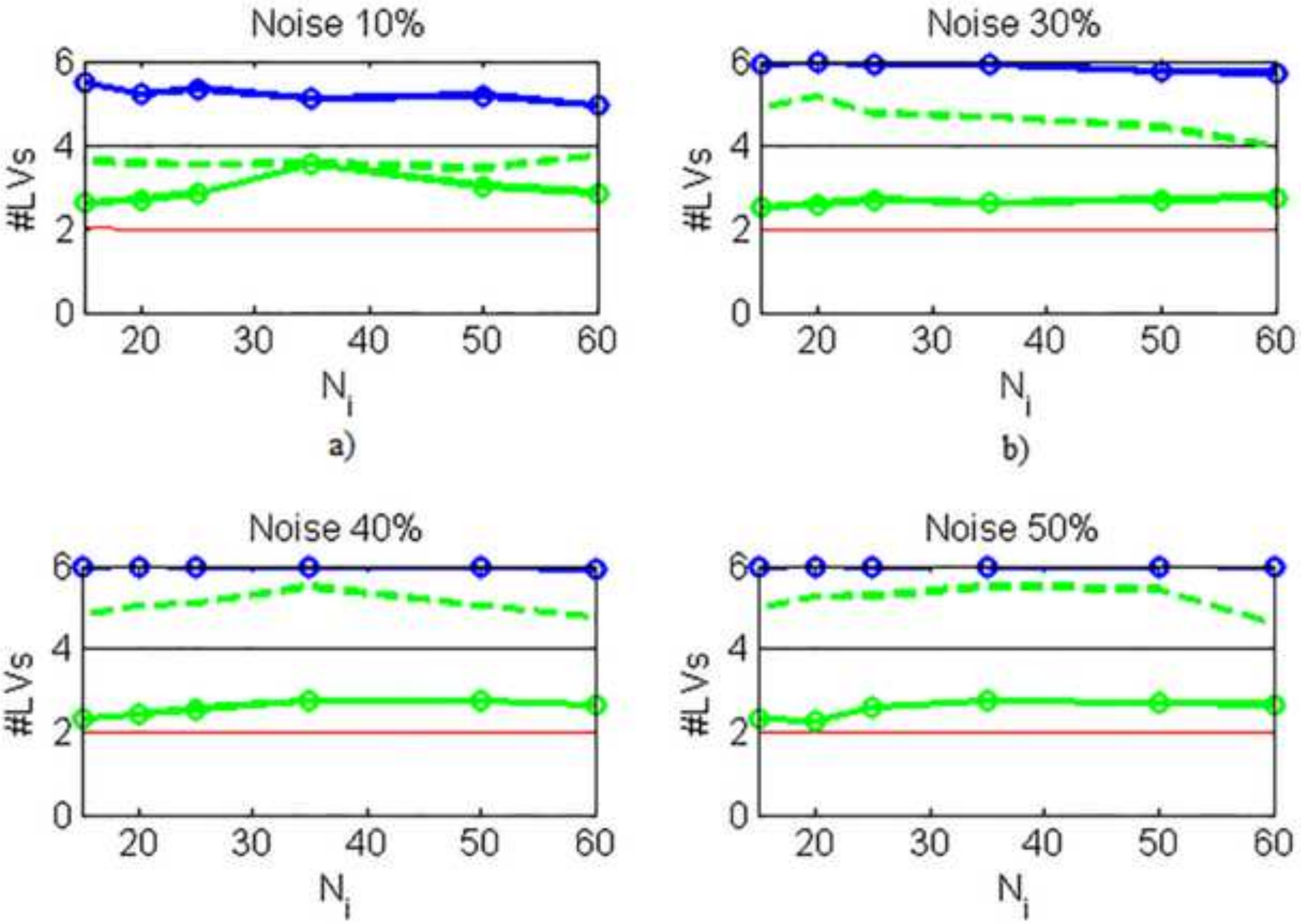
Click here to download high resolution image



Figure(s)

Click here to download high resolution image



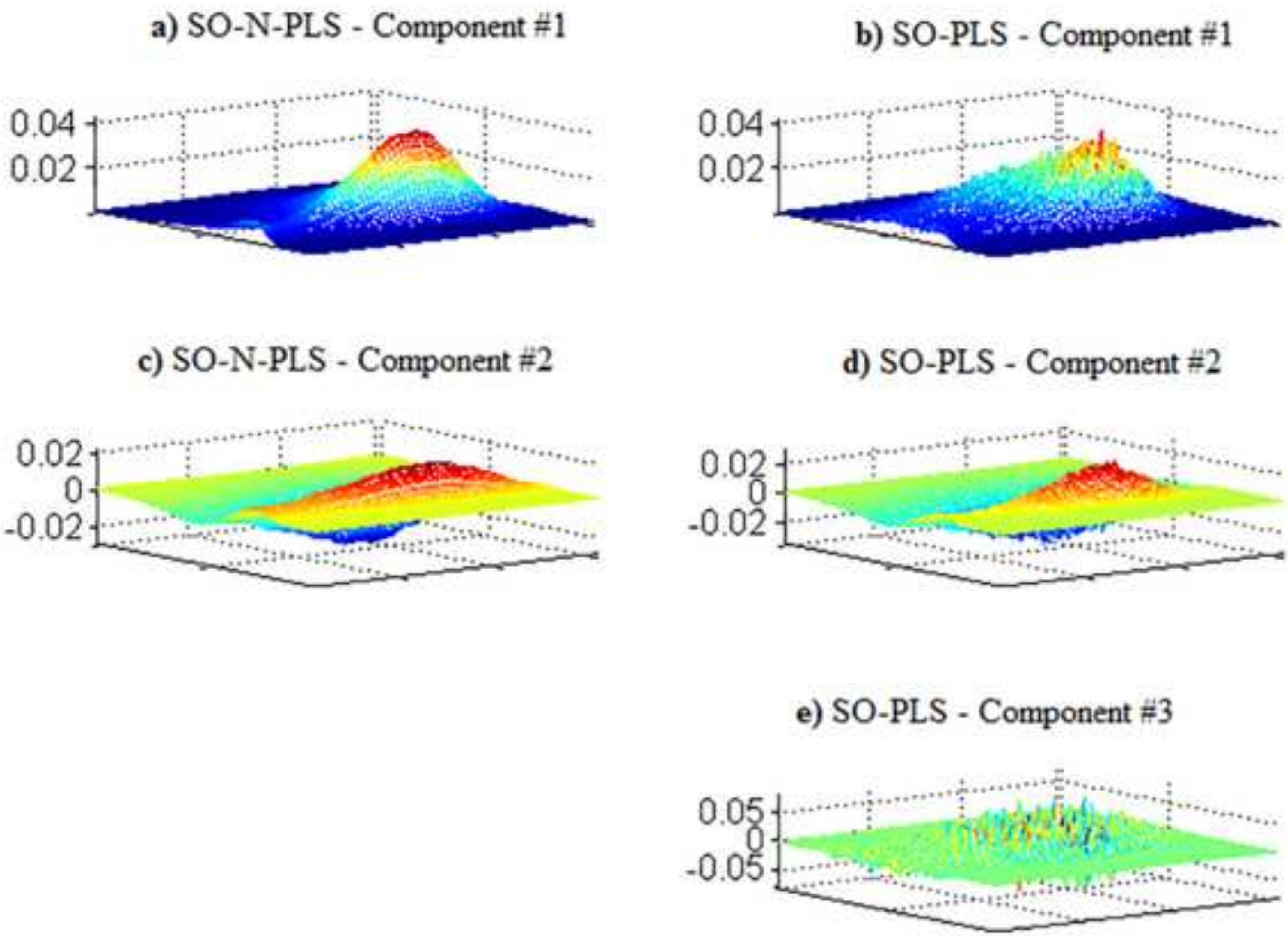


a)

b)

c)

d)



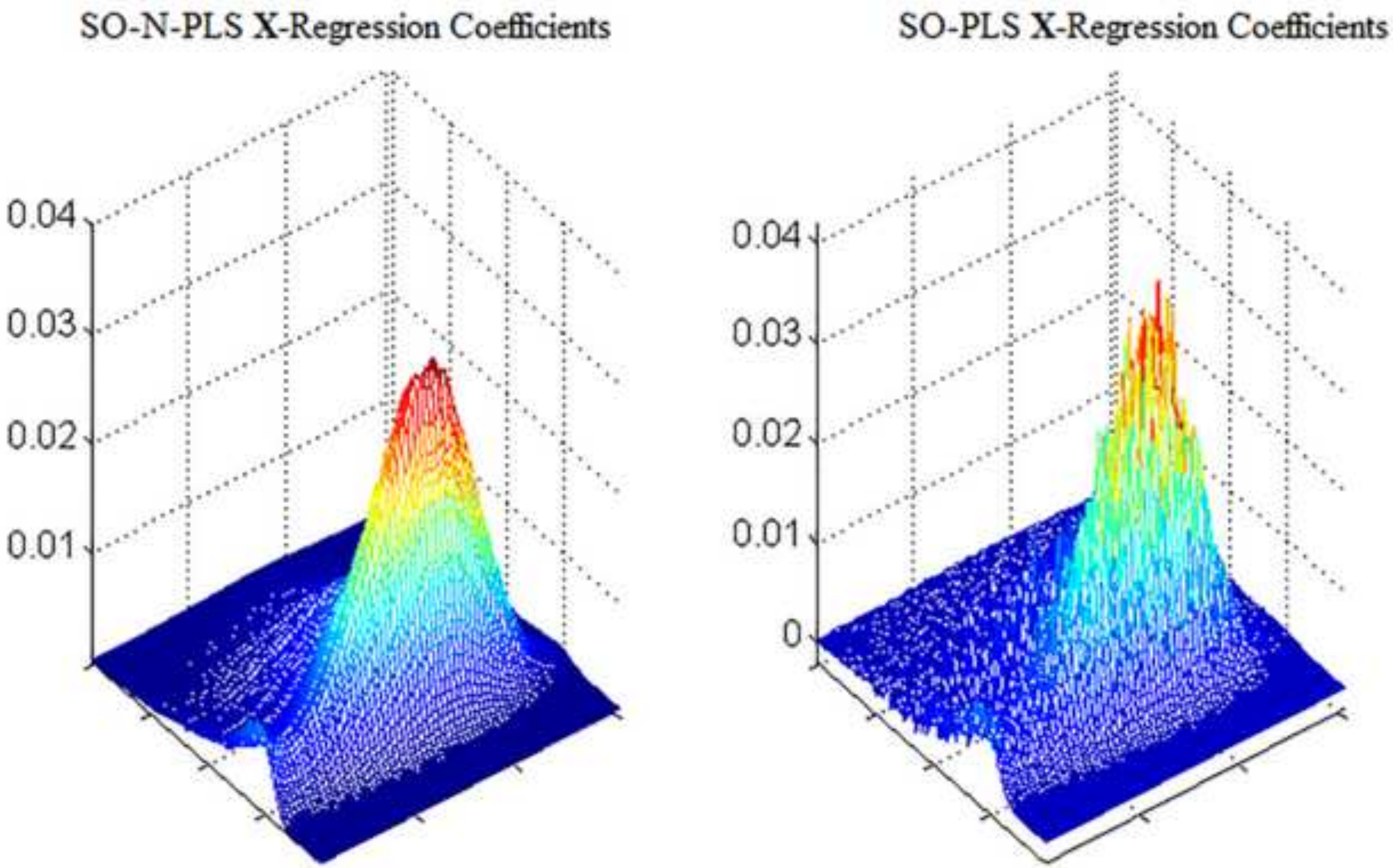
a) SO-N-PLS - Component #1

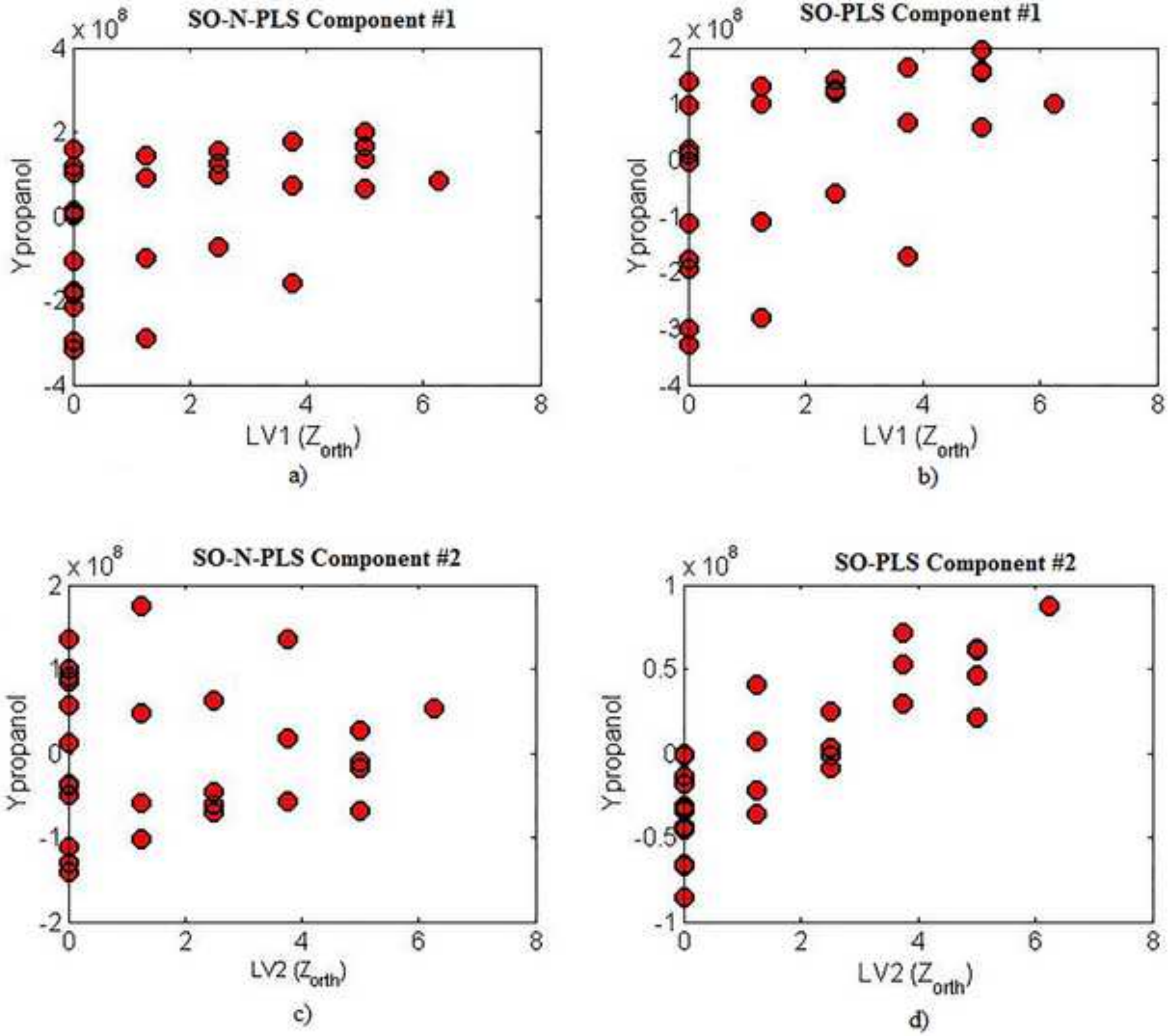
b) SO-PLS - Component #1

c) SO-N-PLS - Component #2

d) SO-PLS - Component #2

e) SO-PLS - Component #3


SO-N-PLS X-Regression Coefficients
0.04
0.03
0.02
0.01
SO-PLS X-Regression Coefficients
0.04
0.03
0.02
0.01
0

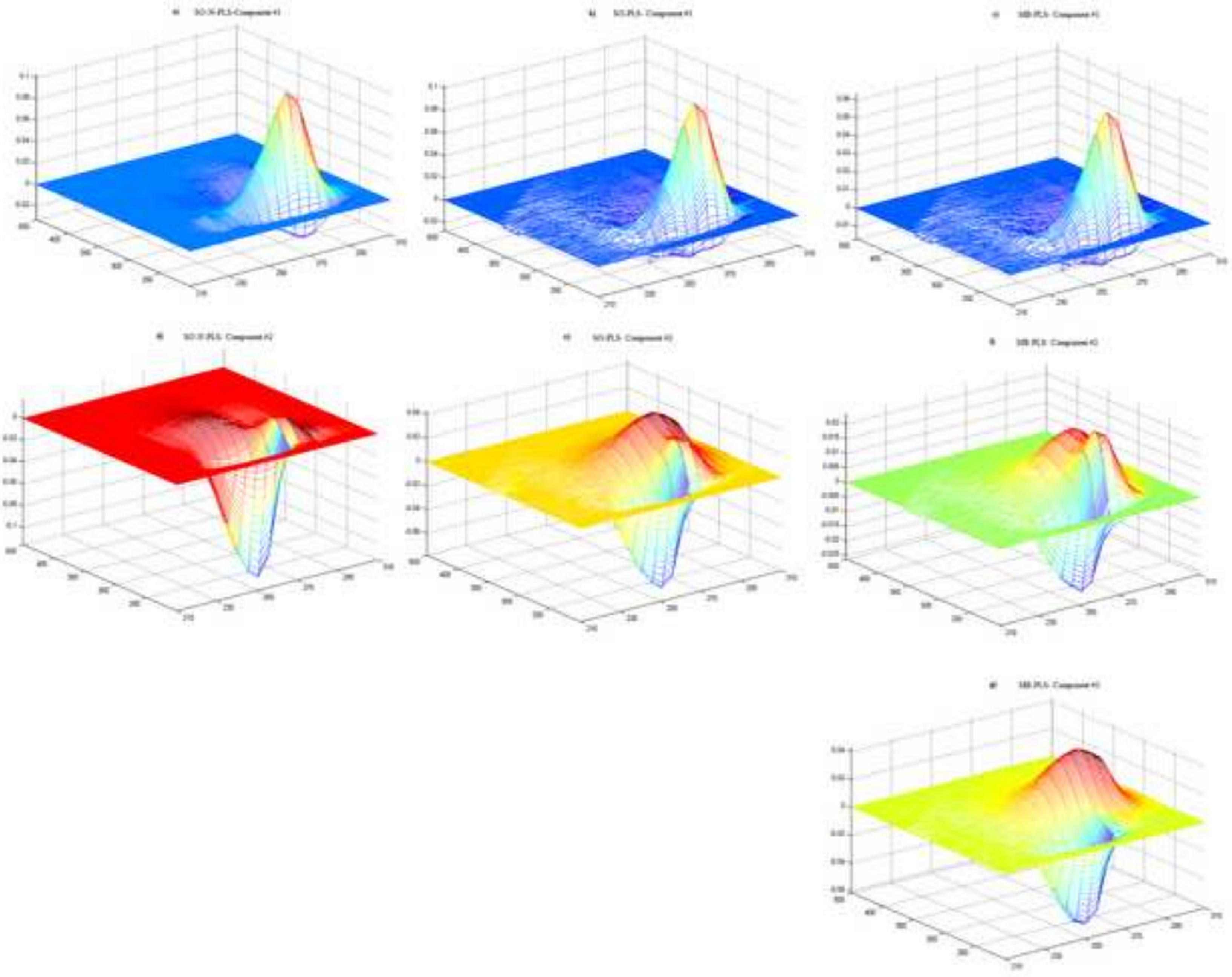

SO-N-PLS Component #1
x 10^8
4
2
0
-2
-4
Ypropanol
0
2
4
6
8
LV1 (Z_orth)
SO-PLS Component #1
x 10^8
2
1
0
-1
-2
-3
-4
Ypropanol
0
2
4
6
8
LV1 (Z_orth)
SO-N-PLS Component #2
x 10^8
2
1
0
-1
-2
Ypropanol
0
2
4
6
8
LV2 (Z_orth)
SO-PLS Component #2
x 10^8
1
0.5
0
-0.5
-1
Ypropanol
0
2
4
6
8
LV2 (Z_orth)

a)

b)

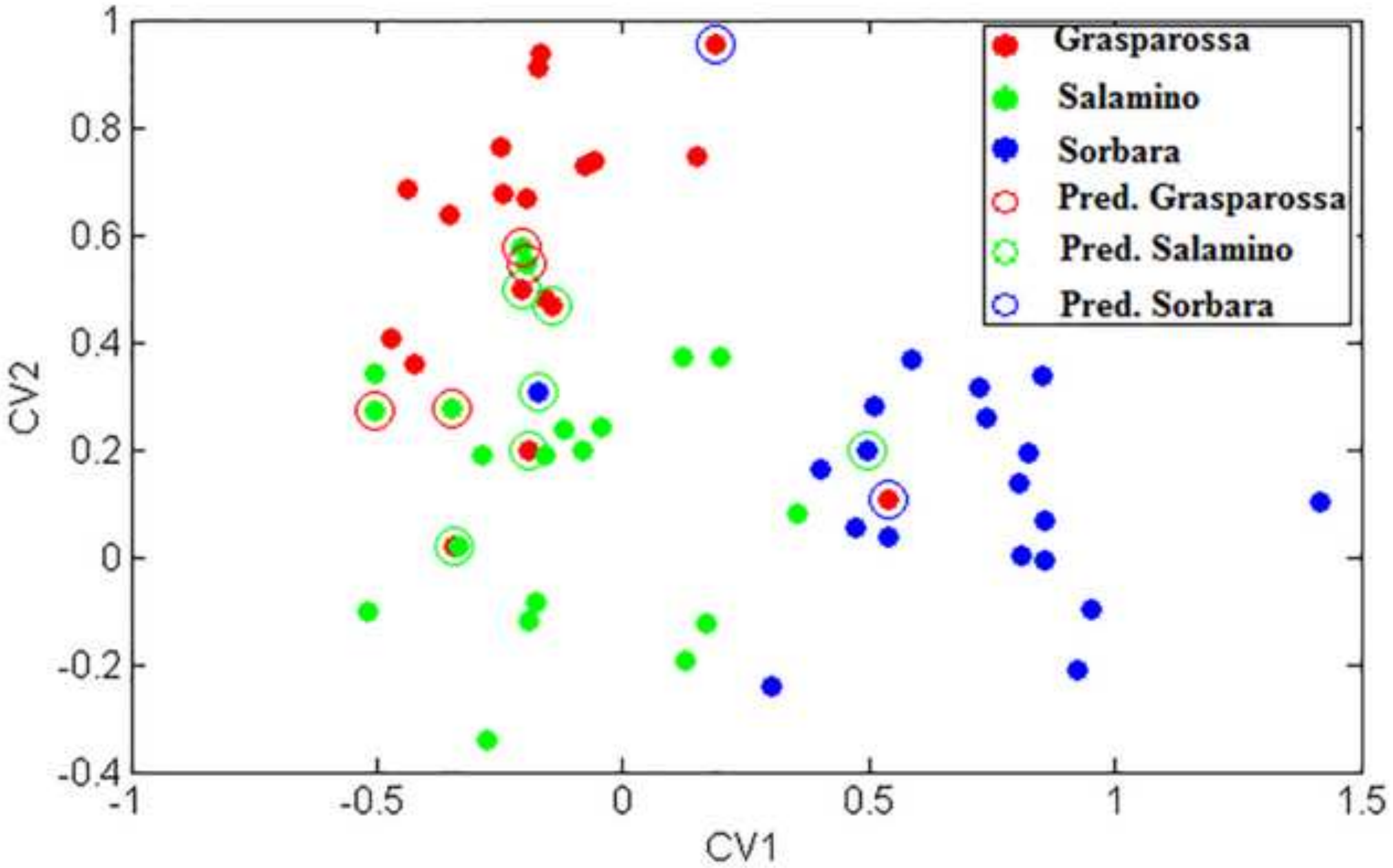
c)

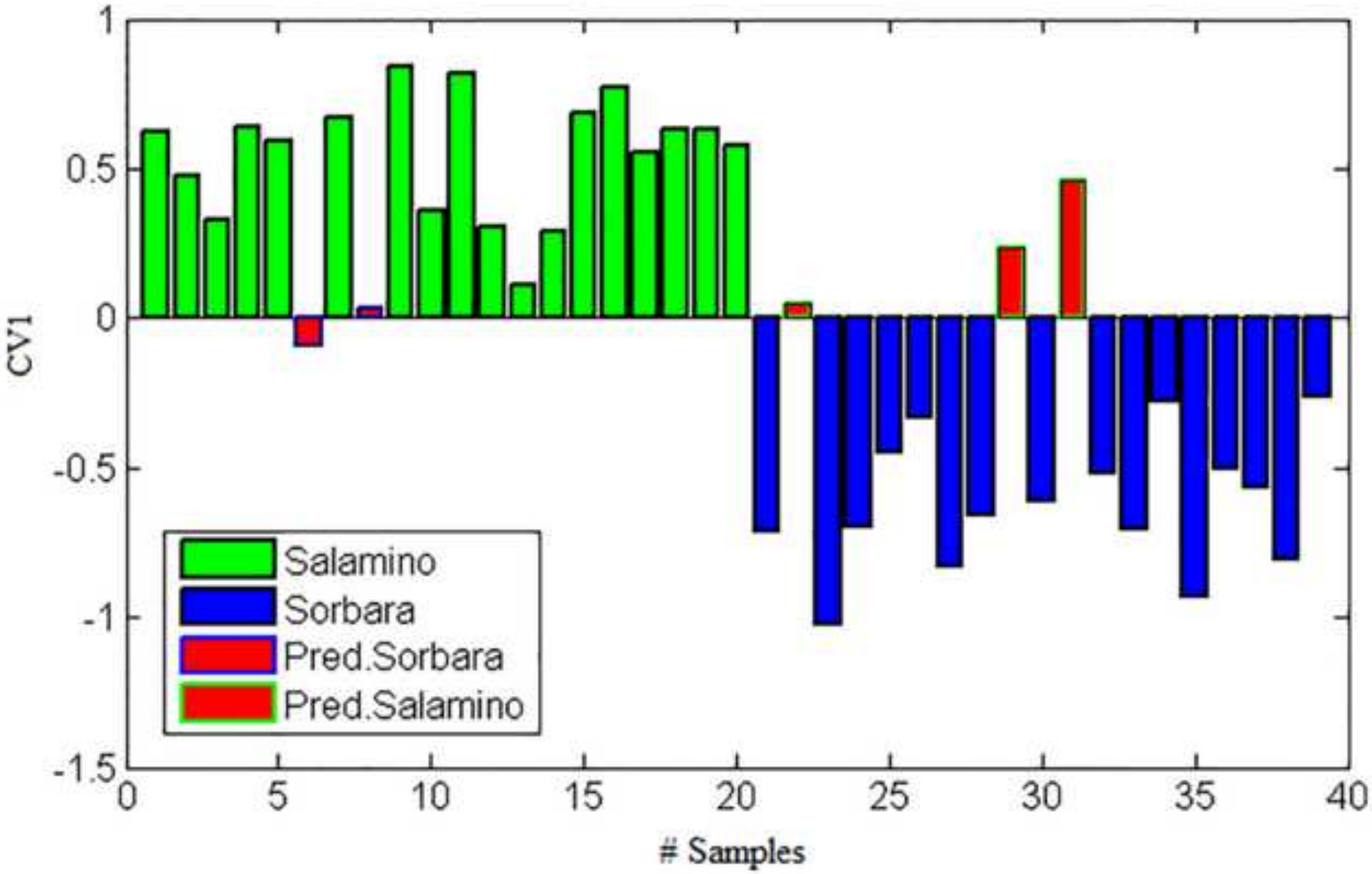
d)


a) SO-N-PLS-Component #1
b) SO-PLS- Component #1
c) MB-PLS- Component #1
d) SO-N-PLS- Component #2
e) SO-PLS- Component #2
f) MB-PLS- Component #2
g) MB-PLS- Component #3

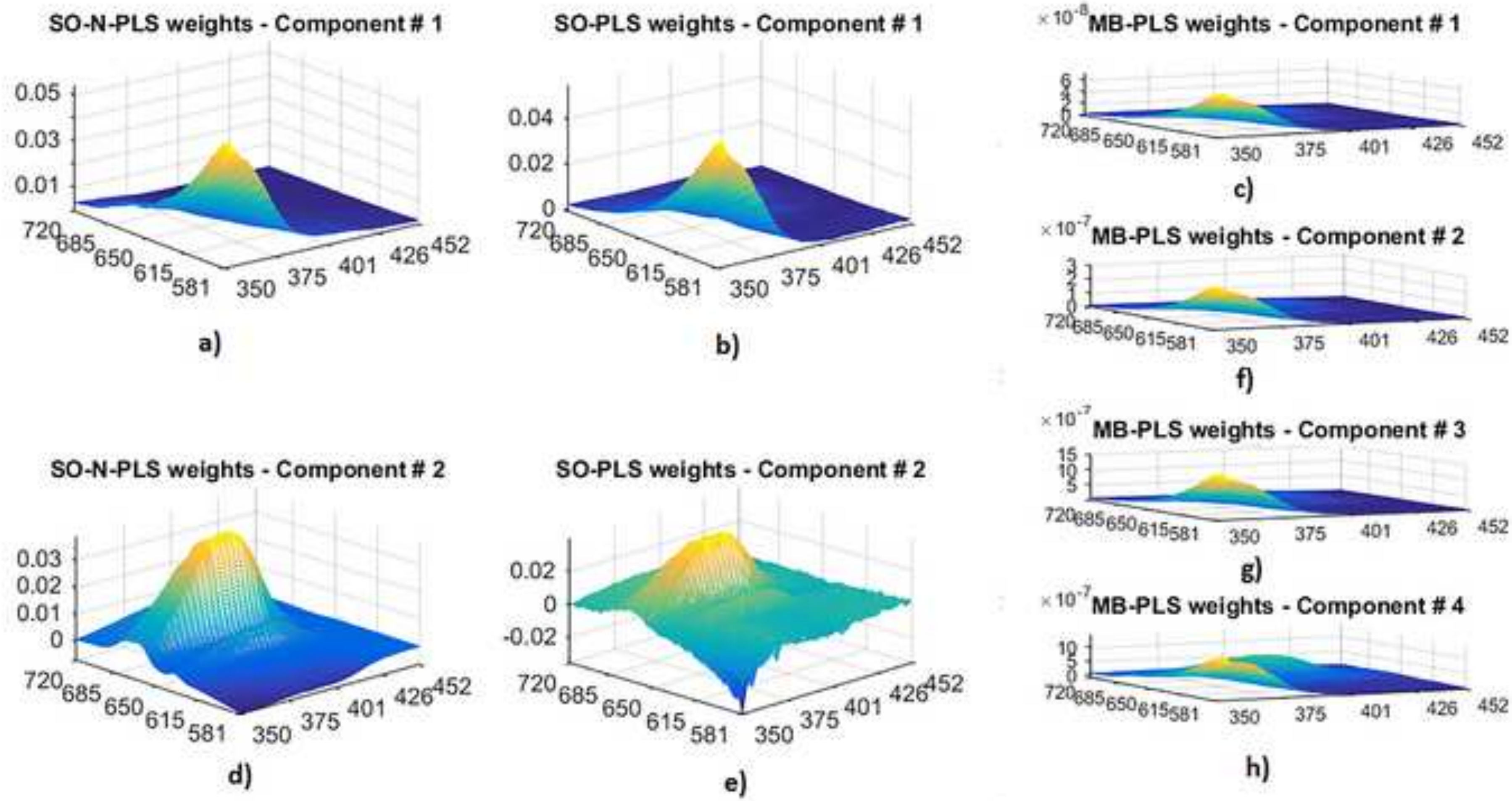
Figure(s)

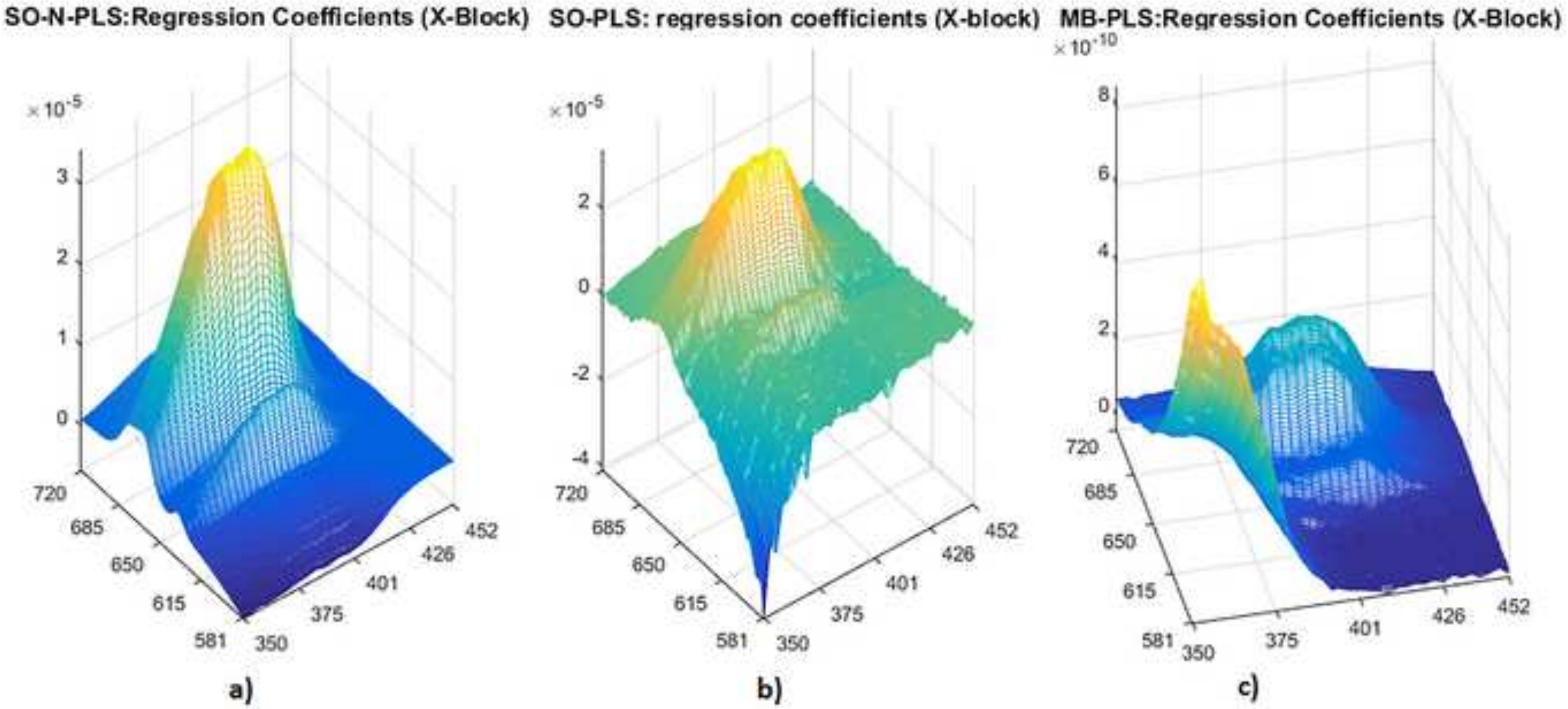
Click here to download high resolution image

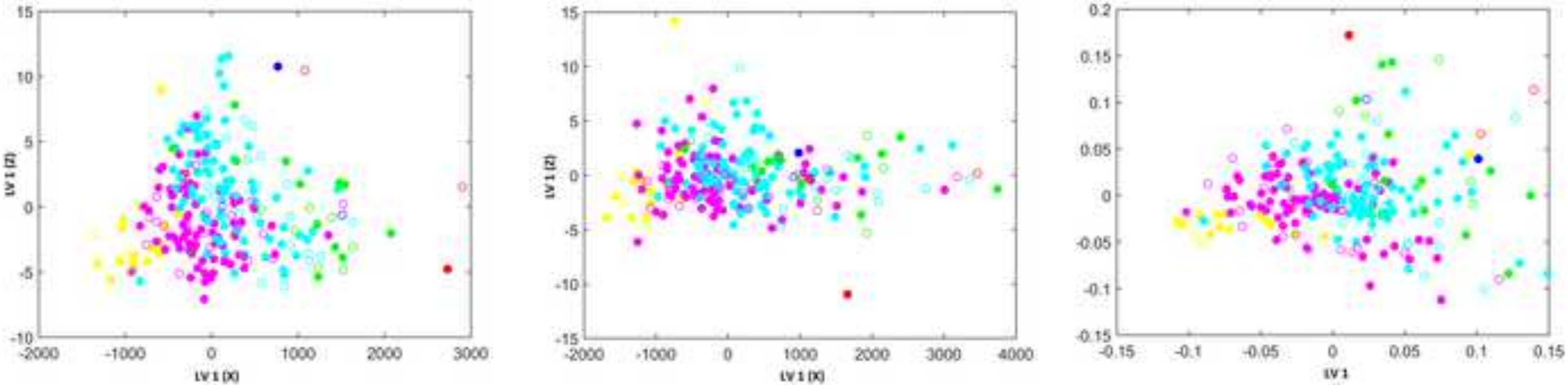



1
0.5
0
-0.5
-1
-1.5
CV1
0
5
10
15
20
25
30
35
40
# Samples
Salamino
Sorbara
Pred.Sorbara
Pred.Salamino

Figure(s)
Click here to download high resolution image




SO-N-PLS:Regression Coefficients (X-Block)
× 10-5
3
2
1
0
720
685
650
615
581
350
375
401
426
452
a)
SO-PLS: regression coefficients (X-block)
× 10-5
2
0
-2
-4
720
685
650
615
581
350
375
401
426
452
b)
MB-PLS:Regression Coefficients (X-Block)
× 10-10
8
6
4
2
0
720
685
650
615
581
350
375
401
426
452
c)


15
10
5
0
-5
-10
-2000
-1000
0
1000
2000
3000
LV 1 (X)
LV 1 (Z)
15
10
5
0
-5
-10
-15
-2000
-1000
0
1000
2000
3000
4000
LV 1 (X)
LV 1 (Z)
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
LV 1

a)

b)

c)