Further investigations on the relationship between the OPLS preprocessing and the NAS

Jean-Claude Boulet, Robert Sabatier

Version of Record: https://www.sciencedirect.com/science/article/pii/S0169743920302616
Manuscript_64f54f3a8d6bad564e1521ea1bd1d50f

Further investigations on the relationship between the O-PLS preprocessing and the NAS.

Jean-Claude Boulet [*†]
Robert Sabatier [‡]

[*]INRAE, UMR1083 Sciences For Enology, SupAgro, University of Montpellier, F-34060 Montpellier, France
[†]Chemhouse Research Group, Montpellier, France
[‡]University of Montpellier, Institut of Functionnal Genomic, CNRS, UMR5203, Inserm U1191, F-34060 Montpellier, France

Abstract

O-PLS is a preprocessing that was presented as an improvement of the PLS algorithm it was issued from. Nevertheless, the bibliography did not confirm, neither for prediction, nor for interpretation. To contribute to a better understanding, we investigated the relationship between O-PLS and the NAS net analyte signal. For four numerical applications, the matrix obtained after the O-PLS deflation tended towards a matrix of rank 1 when the number of removed dimensions increases. Therefore, the line-vectors of this matrix are colinear to the NAS, and so the usual one-latent-variable PLS1 regression after the O-PLS preprocessing can be replaced by almost any regression taking into account from one to all of the variables.

Keywords: PLS, O-PLS, OSC, NAS, prediction, interpretation

1 Introduction

Chemometrics is largely based on the Partial Least Squares or Projection to Latent Structures (PLS) method. Hereafter, all the discussion will apply to the prediction of a single variable. Let $\mathbf{X}$ and $\mathbf{y}$ be a matrix and a column-vector containing the values of variables measured on the same observations. PLS regression was developed to predict the values of $\mathbf{y}$ using $\mathbf{X}$, specially when the variables in $\mathbf{X}$ are highly correlated. For example, PLS regression allowed the prediction of gluten in a wheat sample from a simple near infrared spectrum. Many PLS algorithms have been proposed. Nine of them were compared in Andersson [1]. We proposed a tenth algorithm more focused on the geometry of PLS [2]. Countless applications of PLS can be found in industrial and scientific applications. Nevertheless, in certain fields such as metabolomics, another method is being replacing PLS. This method is based on a preprocessing called Orthogonal Projection to Latent Structures (O-PLS). The O-PLS algorithm was obtained by a slightly modification of the PLS algorithm. To understand it, let us come back to the origin.

The principle of PLS is to extract the subspace in $\mathbf{X}$ that contains the information related to the explanation of $\mathbf{y}$. The idea of the Orthogonal Signal Correction (OSC) preprocessings is to help PLS by removing from $\mathbf{X}$ all information not related to $\mathbf{y}$. OSC are followed by a regression, usually a PLS. A first OSC algorithm, derived from the NIPALS-PLS algorithm, was proposed by Wold et al [3]. This algorithm was latter improved by Trygg and Wold, yielding to a new method called O-PLS [4, 5, 6]. But while other OSC algorithms were proposed [7, 8], the ability of OSC to improve the predictive ability of a regular PLS was being discussed [9, 10]. More attention was paid to O-PLS. It was observed that one O-PLS factor

removed spared one PLS latent variable, without gain in prediction errors [7, 11, 9]. Then Verron et al [12] mathematically confirmed this observation, and Kemsley & Tapp [13] showed a direct relationship between O-PLS and PLS scores. As a conclusion, O-PLS followed by a PLS regression yieds exactly the same prediction than a single PLS regression. Then O-PLS benefits were focused on interpretation rather than prediction, e.g. [14], but even the added value of interpretation was challenged [15], conclusions we agree with. Nevertheless, there is a gap between O-PLS users on the one hand, and mathematical considerations on O-PLS on the other hand. Maybe a geometrical approach could contribute to a better understanding of O-PLS? In particular, Goicoechea et al [10] and Ni et al [16] studied the relationship between OSC methods and the Net Analyte Signal (NAS) introduced by Lorber [17]. We propose to further investigate this property. Notations are presented in the figure 1.

2 Theory.

The name O-PLS let think to a PLS-like method, i.e. a regression method. Certain so-called "O-PLS" functions associate an O-PLS preprocessing followed by a PLS regression with one latent variable. These names are misleading, because O-PLS as introduced in the original papers of Trygg and Wold is clearly a preprocessing, not a calibration method. In order to obtain predictions, O-PLS needs to be followed by a calibration step, e.g. a classical least squares or a PLS regression [10] usually processed with a single latent variable.

O-PLS calculation is iterative. Let A_{max} be a predermined large number of dimensions to be removed. Scores $\mathbf{t_i}$ and loadings $\mathbf{p_i}$ are computed for each value of i between 1 and A_{max}, then the correction is performed

by deflation[9]: $\mathbf{X_{opls}} = \mathbf{X} - \sum_{i=1}^{A_{max}} \mathbf{t_i p_i'}$. Each dimension i is supposed to contain information unrelated to $\mathbf{y}$. But while i is increasing, there will be a breakpoint for which no more information unrelated to $\mathbf{y}$ will be left in $\mathbf{X_{opls}}$. Let A be this value of i; $A \leq A_{max}$. A definition of the NAS is *the part of the signal orthogonal to the other constituents* [17]. This definition was made for spectra, but it can be extended to other signals. Let $\mathbf{s_{nas}}$ be the NAS associated to the compound of interest whose values form $\mathbf{y}$. An interesting property of the NAS is that the product of $\mathbf{X_{opls}}$ by $\mathbf{s_{nas}}$ should be proportional to $\mathbf{y}$. It derives from the following equation [10] : $\mathbf{X_{opls}} = k\mathbf{y}\mathbf{s_{nas}'}$ with k a constant. This situation should occur when an appropriate number of dimensions A has been removed by the OSC. Therefore, the corresponding matrix $\mathbf{X_{opls}}$ should be of rank 1, all the line-vectors of $\mathbf{X_{opls}}$ being collinear. Under these conditions, the classical one-latent-variable PLS1 regression which follows the O-PLS proprocessing could be replaced by *any* other regression without significant loss of predictive ability, even with a simple linear regression. On the other hand, if the theory is not verified, if $\mathbf{X_{opls}}$ is not close enough from a rank-1 matrix, then the predictions should strongly depend on the choice of the regression method. To assess this property, three methods based on O-PLS were compared. The first method noted *OPLS-classic* is the classical combination of an O-PLS preprocessing followed by a PLS regression with 1 latent variable. The second method noted *OPLS-ones* is a modified version of the previous *OPLS-classic*. The modification consisted in replacing the weight vector $\mathbf{w}$ which begins the PLSR algorithm by an arbitrary vector of ones of same dimension. *OPLS-classic* and *OPLS-ones* algorithms are summarized in Table 2. The third method noted *OPLS-univ* is a classical OPLS followed by an univariate regression. The selected variable was chosen as the one with

the largest variability in $\mathbf{X_{opls}}$ after the O-PLS preprocessing removed A_{max} components.

3 Material and methods

Four datasets were selected. Two datasets were challenges proposed by Pierre Dardenne and Vincent Baeten, CRA-Wallonie, at the French annual chemometrics conference in 2007 and 2018. The goal of the 2007 challenge [18] was to predict gluten concentrations. The test dataset was picked up, since it contained a large number of observations (2000). The 2018 challenge consisted in 3908 calibration plus 429 test samples of near infrared spectra. The compound to quantify was not identified. A third dataset was provided by Sylvie Bureau, INRAE, UMR408. Mid infrared spectra ($4000 - 650 cm^{-1}$) were acquired on 750 apricots, for which several reference analysis were performed including refractive index. The last dataset was provided by INRAE, UMR ITAP and SPO. Visible-near infrared spectra were acquired in transmittance and Brix degrees were also measured on 250 grape berries. Processing needed a calibration and a test dataset. Thus, the 2007 challenge, the apricots and grape berries datasets were splitted, the 1500, 600 and 200 observations were assigned to the calibration dataset, the last 500, 150 and 50 observations were assigned to the test dataset, respectively. After centering, the three regression methods described above, *OPLS-classic*, *OPLS-ones* and *OPLS-univ* were processed on the calibration datasets with 1 to A_{max} dimensions removed by OPLS, A_{max} being set to 50 for the apricot datasets and to 40 for the three other datasets. The variable selected for the *OPLS-univ* model was the one with the largest variability after removal of A_{max} dimensions by OPLS. The predictive abilities of the three models were assessed by the calibration errors (RMSEC), the

cross-validation errors (RMSECV) and the prediction errors (RMSEP). The RMSECVs were the average of a random 2-blocks cross-validation repeated 50 times.

4 Results

Results are presented as figures 3, 4, 5 and 6, one for each of the four datasets. After removal of the A_{max} OPLS components, the residual spectra in subfigures (a) present a similar shape, sometimes symmetrical along the axis $y = 0$ which is consistent with a low rank of the corresponding matrix. These subfigures were used to choose the univariate variable of the *OPLS-univ* regression. RMSECs in subfigures (b) are similar for the four datasets. When i is low, a few differences between the three models can be observed. But when i increases, the curves overlay. RMSECV in subfigures (c) also present the same trend for all datasets, but with a different behavior according to the models. The RMSECV of *OPLS-univ* decreases steadly when i increases. On the other hand, *OPLS-classic* and *OPLS-ones* present the classical shape of a decrease, then an increase. When i is high, those two curves overlay. Then, RMSEP in subfigures (d) can present differences between models. But when i is high, the curves of the three models overlay.

5 Discussion and conclusion

The hypothesis that, by increasing the number i of removed dimensions, the rank of the resulting matrix $\mathbf{X_{opls}}$ gets close to 1, still stands up after the benchmark on 4 datasets. Similar results, not presented, confirmed this observation. When i increases, all the line vectors of $\mathbf{X_{opls}}$ tend to be colinear. This situation occured when i was higher than $A = 7$, 15, 40 and 5 for the

2007 and 2018 challenges, the apricots and grape berries datasets respectively. These values of A need to be compared to A_{opt} the optimal number of dimensions to be removed, determined by the RMSECV curves, around $13-20$, $15-20$, $7-12$ and $15-22$ respectively. For the apricot dataset, A_{opt} corresponds to a situation where the three models *OPLS-classic*, *OPLS-ones* and *OPLS-univ* do not yield the same prediction, and where prediction errors remain high. Figure 5(d) suggests that a higher number of dimensions, A_{opt} around $20-22$, would have yielded better predictions, but such values remain before the convergence region obtained for $A = 40$ ($A > A_{opt}$). For the three other datasets, convergence of the three models had already been achieved: $A < A_{opt}$, and the models were found more robust when applied to the test datasets. Maybe a clue to identify robust ($A < A_{opt}$) from not robust ($A > A_{opt}$) models?

Finally, all the information about the NAS is supported by the $\mathbf{X}_{opls}$ matrix, issued from the O-PLS preprocessing. The weights $\mathbf{w}$ and the regression vector $\mathbf{b}$ of the PLSR in the *OPLS-classic* method are linear combination of the lines of $\mathbf{X}_{opls}$, weighed by the $\mathbf{y}$ values. Setting $\mathbf{w}$ to a vector of ones in *OPLS-ones* is another way to weight the lines of $\mathbf{X}_{opls}$, each line has the same weight. When $\mathbf{X}_{opls}$ is close to a rank-1 matrix, these calculations yield the same result. The $\mathbf{X}_{osc}$ matrix or the regression coefficients $\mathbf{b}$ can all be interpreted as an estimation of the NAS. According to its definition, the NAS is obtained after an orthogonal (or oblique) projection, and therefore inherits of the properties of such calculations. In particular, the NAS is expected to present features from the compound of interest (whose concentrations are $\mathbf{y}$), but more ennoying it can present features from the other coexisting constituents [19], and eventually it can have dropped features that were present in the compound of interest. So the interpretation

of the NAS remains tricky and should be confirmed by other informations. These considerations also apply to the residual matrix, $\mathbf{X} - \mathbf{X}_{opls}$, which is the result of an orthogonal projection, too. Note that the same interpretation of the NAS calculated from O-PLS can also be performed with the regression coefficients of a PLS regression, already reported as an estimation of the NAS [20].

6 Acknowledgements

Thanks to Vincent Baeten, Pierre Dardenne and Cécile Barron, for providing datasets; and Christelle Reynès and JM Roger for useful comments on the manuscript.

References

[1] Martin Andersson. A comparison of nine pls1 algorithms. *Journal of Chemometrics*, 23:518–529, 2009.

[2] Jean-Claude Boulet, Dominique Bertrand, Gérard Mazerolles, Robert Sabatier, and Jean-Michel Roger. A family of regression methods derived from standard plsr. *Chemometrics and Intelligent Laboratory Systems*, 120:13–25, 2013.

[3] Svante Wold, Henrik Antti, Frederik Lindgren, and Jerker Ohman. Orthogonal signal correction of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44:175–185, 1998.

[4] J.Trygg. *Parsimonious multivariate models*. PhD thesis, Umea University, Sweden, 2001.

[5] Yohan Trygg and Svante Wold. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16:119–128, 2002.

[6] Johan Trygg and Svante Wold. Orthogonal signal projection. Patent US20030200040A1, 2003.

[7] Tom Fearn. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 50:47–52, 2000.

[8] Johan A. Westerhuis, Sijmen De Jong, and Age K. Smilde. Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 23:13–25, 2001.

[9] O. Svensson, T. Kourti, and J.F. MacGregor. An investigation of orthogonal signal correction algorithms and their characteristics. *Journal of Chemometrics*, 16:176–188, 2002.

[10] Hector C. Goicoechea and Alejandro C. Olivieri. A comparison of orthogonal signal correction and net analyte preprocessing methods. theoretical and experimental study. *Chemometrics and Intelligent Laboratory Systems*, 56:73–81, 2001.

[11] Agnar Hoskuldsson. Variable and subset selection in pls regression. *Chemometrics and Intelligent Laboratory Systems*, 55:23–38, 2001.

[12] Thomas Verron, Robert Sabatier, and Richard Joffre. Some theoretical properties of the o-pls method. *Journal of Chemometrics*, 18:62–68, 2004.

[13] E.K. Kemsley and H.S. Tapp. Opls filtered data can be obtained directly from non-orthogonalized pls1. *Journal of Chemometrics*, 23:263–264, 2009.

[14] Matthias Hedenstrom, Susanne Wiklund, Bjorn Sundberg, and Ulf Edlund. Visualization and interpretation of o-pls models based on 2d nmr data. *Chemometrics and Intelligent Laboratory Systems*, 92:110–117, 2008.

[15] Ulf G. Indahl. The o-pls methodology for orthogonal signal correction - is it correcting or confusing. *Journal of Chemometrics*, pages 1–14, 2017.

[16] Wangdong Ni, Steven.D. Brown, and Ruilin Man. The relationship between net analyte signal/preprocessing and orthogonal signal correction algorithms. *Chemometrics and Intelligent Laboratory Systems*, 98:97–107, 2009.

[17] Avraham Lorber, Klass Faber, and Bruce R. Kowalski. Net analyte signal calculation in multivariate calibration. *Analytical Chemistry*, 69:1620–1626, 1997.

[18] J.A. Fernandez Pierna, F. Chauchard, S. Preys, J.M. Roger, O. Galtier, V. Baeten, and P. Dardenne. How to build a robust model against perturbation factors with only a few reference values: a chemometric challenge at chimiometrie 2007. *Chemometrics and Intelligent Laboratory Systems*, 106:152–159, 2011.

[19] J.C. Boulet and J.M. Roger. Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, 117:61–69, 2012.

[20] Joan Ferré and Nicolaas M. Faber. Net analyte signal calculation for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 69:123–136, 2003.

7 Captions

Figure 1: Notations

Figure 2: *OPLS-classic* a regular O-PLS followed by a PLS1 with one latent variable (left); *OPLS-ones* a regular O-PLS followed by a w-modified PLS1, **w** being set to a vector of ones (right)

Figure 3: Chimiometrie 2007 challenge, NIR spectra: (a) spectra after 40 OPLS components removed, the vertical line representing the variable selected for *OPLS-univ*; (b-d) RMSEC, RMSECV and RMSEP for the 3 models: *OPLS-classic* (red), *OPLS-ones* (green) and *OPLS-univ* (blue)

Figure 4: Chimiometrie 2018 challenge, NIR spectra: (a) spectra after 40 OPLS components removed, the vertical line representing the variable selected for *OPLS-univ*; (b-d) RMSEC, RMSECV and RMSEP for the 3 models: *OPLS-classic* (red), *OPLS-ones* (green) and *OPLS-univ* (blue)

Figure 5: Apricots MIR spectra: (a) spectra after 50 OPLS components removed, the vertical line representing the variable selected for *OPLS-univ*; (b-d) RMSEC, RMSECV and RMSEP for the 3 models: *OPLS-classic* (red), *OPLS-ones* (green) and *OPLS-univ* (blue)

Figure 6: Grape fruit NIR reflectances: (a) spectra after 40 OPLS components removed, the vertical line representing the variable selected for *OPLS-univ*; (b-d) RMSEC, RMSECV and RMSEP for the 3 models: *OPLS-classic*

(red), *OPLS-ones* (green) and *OPLS-univ* (blue)

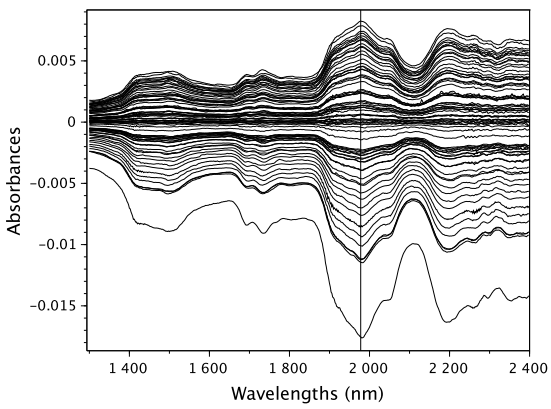
8 Figures

N	number of observations
Q	number of variables
i	number of O-PLS components removed
A	minimum value of i for which $\mathbf{X}_{osc}$ can be approximated to a rank-1matrix
A_{opt}	value or range of values of i for which the RMSECV is minimum
A_{max}	a value of i larger than A and A_{opt}
$\mathbf{X}$	matrix of N lines and Q columns
$\mathbf{y}$	column vector of N elements
$\mathbf{X}_{osc}$	the matrix $\mathbf{X}$ after OSC correction
$\mathbf{X}_{opls}$	the matrix $\mathbf{X}$ after O-PLS correction
$\mathbf{t}_i$	i^{th} score vector for OSC/OPLS
$\mathbf{p}_i$	i^{th} loading vector for OSC/OPLS
$\mathbf{w}$	weight vector of PLS
$\mathbf{t}$	score vector of PLS
$\mathbf{b}$	regression vector of PLS
$\mathbf{1}_Q$	a vector of ones of length Q
$\mathbf{s}_{nas}$	the NAS net analyte signal

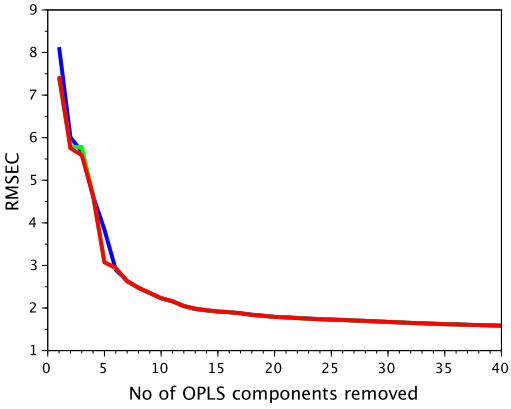
Figure 1:

$\mathbf{X_{opls}}$ calculated with O-PLS	
$\mathbf{w} = \mathbf{X'_{opls}y}$	$\mathbf{w} = \mathbf{1_Q}$
$\mathbf{t} = \mathbf{X_{opls}w}$	$\mathbf{t} = \mathbf{X_{opls}1_Q}$
$c = \mathbf{y't}(\mathbf{t't})^{-1}$	$c = \mathbf{y't}(\mathbf{t't})^{-1}$
$\mathbf{b} = c\mathbf{w}$	$\mathbf{b} = c\mathbf{1_Q}$

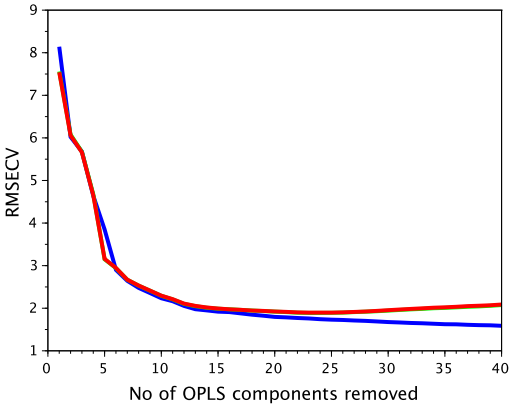
Figure 2:



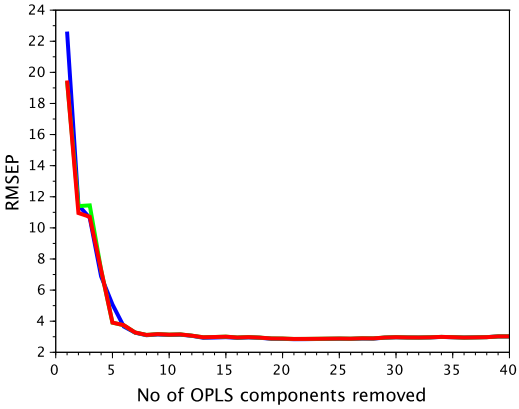
(a)



(b)

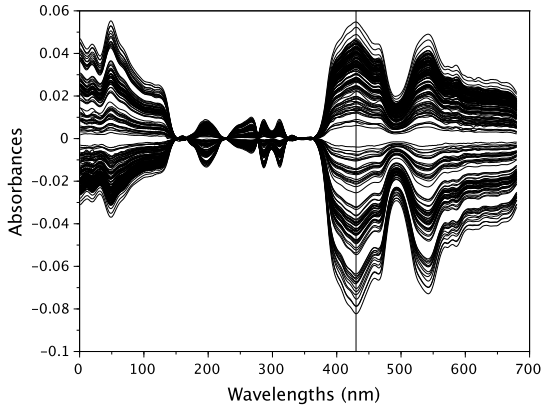


(c)

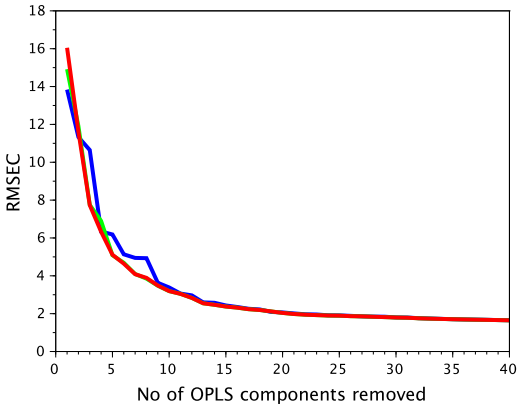


(d)

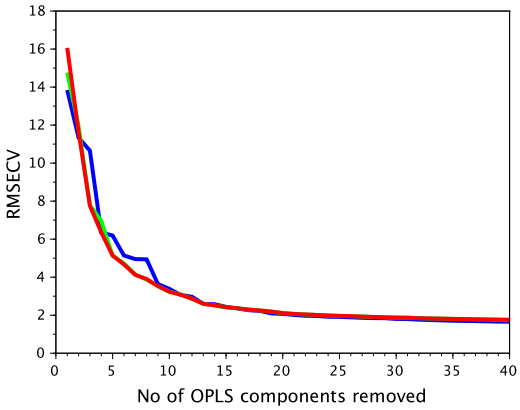
Figure 3:



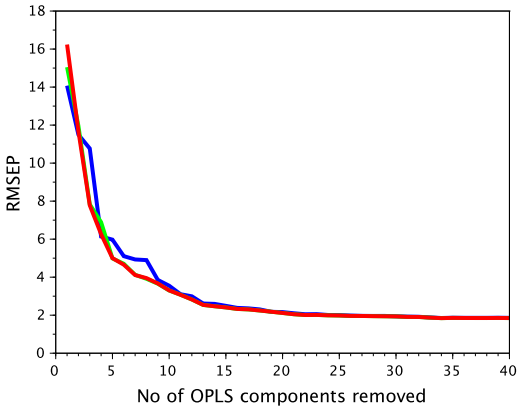
(a)



(b)

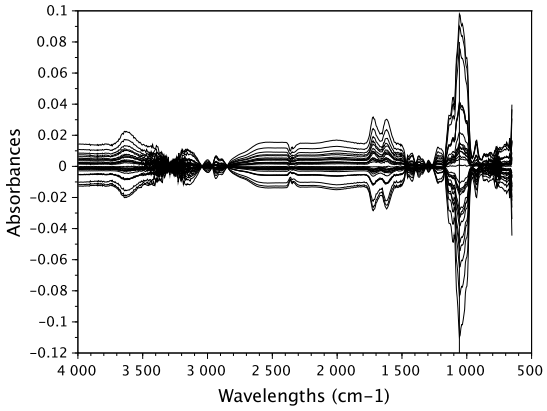


(c)

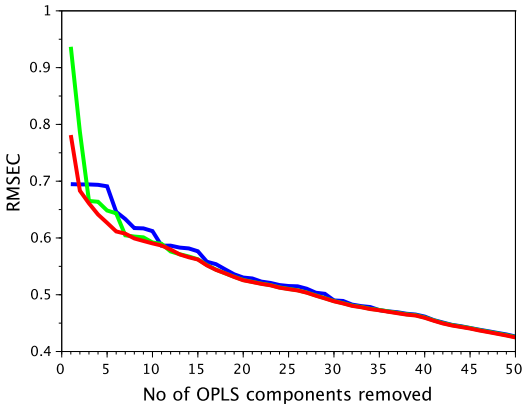


(d)

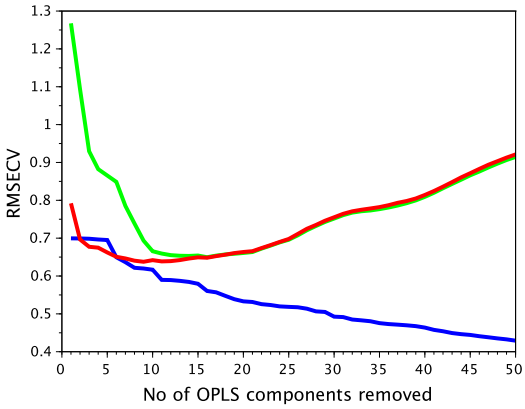
Figure 4:



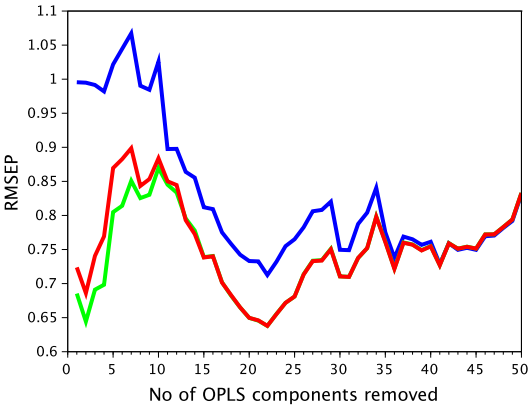
(a)



(b)

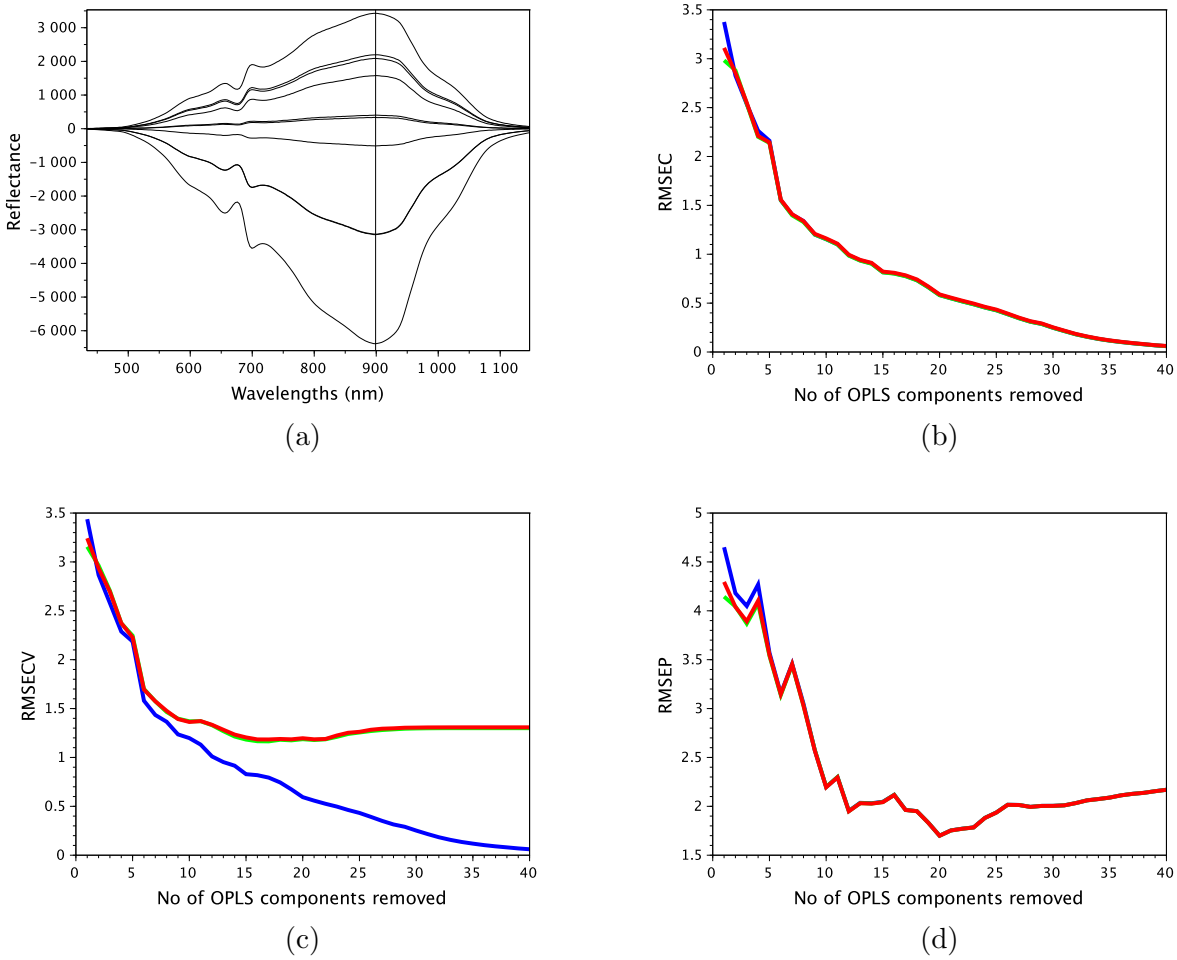


(c)



(d)

Figure 5:



(a) (b) (c) (d)

Figure 6: