Received: 12 November 2013, Revised: 4 April 2014, Accepted: 10 April 2014, Published online in Wiley Online Library: 15 May 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2627

Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS)

Beatriz Galindo-Prieto[a,b], Lennart Eriksson[c] and Johan Trygg[a]*

A new approach for variable influence on projection (VIP) is described, which takes full advantage of the orthogonal projections to latent structures (OPLS) model formalism for enhanced model interpretability. This means that it will include not only the predictive components in OPLS but also the orthogonal components. Four variants of variable influence on projection (VIP) adapted to OPLS have been developed, tested and compared using three different data sets, one synthetic with known properties and two real-world cases. Copyright © 2014 John Wiley & Sons, Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: variable influence on projection; VIP; OPLS; variable selection; PLS

1. INTRODUCTION

Multivariate analysis based on partial least squares (i.e., PLS and orthogonal projections to latent structures [OPLS]) models has become a useful and appreciated toolbox in research and industrial environments. Projections to latent structures by means of PLS was described by Herman and Svante Wold in 1983 [1], Geladi and Kowalski in 1986 [2], Wold and Cocchi in 1993 [3], and Eriksson and Wold in 2001 [4]; and OPLS was presented by Trygg and Wold in 2002 [5]. OPLS separates the systematic variation in X into two parts, one that is correlated (predictive) to Y and another part that is uncorrelated (orthogonal) to Y.

It has been shown that predictions by PLS and orthogonal methods for single-y problems perform equally well provided that identical model complexity and cross-validations are compared [6,7]. Nonetheless, model interpretation may differ between OPLS and PLS as the predictive and orthogonal variations are highlighted by OPLS. This in turn may improve decision-making, that is, OPLS can perform better than PLS [8,9]. For example, the model interpretation and multiple subjective decisions are important for constructing a valid prediction model. This includes selection of observations, variables, scaling, preprocessing methods, transformations, and quality control as demonstrated by Shi *et al.* in a large international study coordinated by the Food and Drug Administration and published in Nature Biotechnology [10].

For PLS, variable influence on projection (VIP) is an established parameter that summarizes the importance of the X-variables in a PLS model with many components [3,11]. Other useful model parameters include, for instance, loading weights or regression coefficients [12–14]. What matters is not to find a single parameter to interpret, for example, VIP, but to synergistically use VIP, weights, regression coefficients, and loadings to assess the variables, and their contribution to the model.

The selection of which parameter to use depends on the configuration of the data set and the complexity of the model. Interestingly, the most compact model interpretation alternative is offered by VIP. For a given PLS model and data set, there can always only be one single VIP expression, regardless of the number of components in the model and the number of responses in the Y-matrix. This parsimony and its intuitive interpretation promote the popularity of the VIP parameter [15–19]. In this work, we aimed to reformulate VIP to take full advantage of the OPLS model formalism for enhanced model interpretability. The VIP for OPLS should include not only the predictive components but also the orthogonal components.

2. THEORY

2.1. Variable influence on projection applied to partial least squares

VIP is a parameter used for calculating the cumulative measure of the influence of individual X-variables on the model [20]. For a given PLS dimension, a, the squared PLS weight $(W_a)^2$ of that term is multiplied by the explained sum of squares (SS) of that PLS dimension; and the value obtained is then divided by the total explained SS by the PLS model and multiplied by the number of terms in the model. The final VIP is the square root of that number. Equation 1 gives a detailed view of the VIP calculation.

$$VIP_{PLS} = \sqrt{K \times \left(\frac{\left[\sum_{a=1}^{A} \left(W_a^2 \times SSY_{comp,a} \right) \right]}{SSY_{cum}} \right)} \qquad (1)$$

* *Correspondence to: Johan Trygg, Computational Life Science Cluster (CLiC), Department of Chemistry, Umeå University, Umeå, Sweden.*
E-mail: johan.trygg@chem.umu.se

[a] *B. Galindo-Prieto, J. Trygg*
Computational Life Science Cluster (CLiC), Department of Chemistry, Umeå University, Umeå, Sweden

[b] *B. Galindo-Prieto*
Industrial Doctoral School (IDS), Umeå University, Umeå, Sweden

[c] *L. Eriksson*
MKS Umetrics, Umeå, Sweden

According to Equation 1, VIP is a weighted combination over all components of the squared PLS weights (W_a), where $SSY_{comp,a}$ is the sum of squares of Y explained by component a, A is the total number of components, and K is the total number of variables. The average VIP is equal to 1 because the SS of all VIP values is equal to the number of variables in X. This means that if all X-variables have the same contribution to the model, they will have a VIP value equal to 1. VIP values larger than 1 point to the most relevant variables, and generally VIP values below 0.5 are considered irrelevant variables.

2.2. Variable influence on projection for orthogonal projections to latent structures

The VIP concept cannot be directly transferred from PLS to OPLS. The reason for this can be understood by considering the expression of VIP seen in Equation 1. The weighting of the squared *w*-values that takes place is based on the explained sum of squares of Y (SSY). This weighting is sensible for the predictive component in OPLS, which will have an explained SSY different from zero, but not applicable to any occurring orthogonal component because the latter by definition does not explain any systematic structure of Y (hence, the SSY will be zero). Consequently, the use of SSY only corresponds to an ordering of variables equivalent to the predictive component loading. In order to explore the variable influences of the full OPLS model, the contribution from orthogonal components should be included. Therefore, there is a need to adapt the classical PLS-VIP expression such that it better applies to OPLS. This includes the use of not only SSY (amount of variation in Y explained by the model) but also the use of SSX (amount of variation in X explained by the model). Loading weights (*w*) are used for VIP calculation in PLS models, but for OPLS, we also introduce the use of the normalized loadings (*p*), in analogy with the normalized loading weights (*w*). This results in four different variants of VIP for OPLS that we include to be evaluated (Table I). It should be observed that for the OPLS predictive component, *w* and *p* loadings will be very similar [21] but not for the OPLS orthogonal components.

The four variants of VIP (VIP 1–4) are described in the following text. $VIP_{i,o}$ stands for the value of the VIP variant *i* for the orthogonal components, $VIP_{i,p}$ corresponds to the value of the VIP variant *i* for the predictive components, and $VIP_{i,tot}$ represents the total sum for both predictive and orthogonal parts of the OPLS model for VIP variant *i*. Predictive components will be represented by *a*, and orthogonal components will be represented by a_o. Analogously, A_p is the total number of predictive components, and A_o is the total number of orthogonal components. *K* is the total number of variables (Equations 9 and 11 describe *K* for orthogonal and predictive components, respectively). The SS has the subscript *comp,a* for the explained SS of a^{th} component, the subscript *comp,a*$_o$ for the explained SS of a_o^{th} component, and the subscript *cum* for the cumulative (i.e., total) explained SS by all components in the model.

Table I. The four VIP for OPLS variants

OPLS-VIP variant	Loadings	Weighting parameter for predictive components	Weighting parameter for orthogonal components
VIP_1	W	SSY	SSX
VIP_2	P	SSY	SSX
VIP_3	W	[SSY,SSX]	[SSY,SSX]
VIP_4	P	[SSY,SSX]	[SSY,SSX]

OPLS, orthogonal projections to latent structures; VIP, variable influence on projection.

VIP_1 is the first variant, which is calculated based on loading weights (*w*) using SSY for the predictive component and SSX for the orthogonal component. $VIP_{1,o}$ corresponds to the value of VIP for the orthogonal components using SSX (Equation 2), and $VIP_{1,p}$ corresponds to the value of VIP for the predictive components using SSY (Equation 3); the VIP value for both predictive and orthogonal components is calculated according to Equation 4.

$$VIP_{1,o} = \sqrt{K \times \left(\frac{\left[\sum_{a_o=1}^{A_o}\left(Wo_{a_o}^2 \times SSX_{comp,ao}\right)\right]}{SSX_{cum,o}} \right)} \qquad (2)$$

$$VIP_{1,p} = \sqrt{K \times \left(\frac{\left[\sum_{a=1}^{A_p}\left(W_a^2 \times SSY_{comp,a}\right)\right]}{SSY_{cum,p}} \right)} \qquad (3)$$

$$VIP_{1,tot} = \sqrt{\frac{1}{2} \times \left(VIP_{1,o}^2 + VIP_{1,p}^2\right)} \qquad (4)$$

The second variant (**VIP_2**; Equation 5–7) is similar to **VIP_1** but now using the normalized loadings (*p*).

$$VIP_{2,o} = \sqrt{K \times \left(\frac{\left[\sum_{a_o=1}^{A_o}\left(Po_{a_o}^2 \times SSX_{comp,ao}\right)\right]}{SSX_{cum,o}} \right)} \qquad (5)$$

$$VIP_{2,p} = \sqrt{K \times \left(\frac{\left[\sum_{a=1}^{A_p}\left(P_a^2 \times SSY_{comp,a}\right)\right]}{SSY_{cum,p}} \right)} \qquad (6)$$

$$VIP_{2,tot} = \sqrt{\frac{1}{2} \times \left(VIP_{2,o}^2 + VIP_{2,p}^2\right)} \qquad (7)$$

The combination [SSY, SSX] is introduced in the third and fourth variants (VIP_3 and VIP_4). VIP_3 is calculated by means of loading weights (*w*), while VIP_4 is based on normalized loadings (*p*). These equations have been written for the general multi-y case, which implies that $SSY_{comp,ao}$ (which is computed as the difference between SSY_{ao-1} and SSY_{ao}) can have a value different to zero. The key to this is the number of predictive components and the number of variables in Y. If the number of predictive components is equal to the number of variables in Y, then the value of $SSY_{comp,ao}$ will be zero; but if the number of predictive components is less than the number of variables in Y, then the orthogonal components can, in fact, have some correlation to Y, and $SSY_{comp,ao}$ will be small, close to zero, but not strictly zero. For single-y cases, this value will always be zero.

As in the previous two variants, three equations describe the three VIP vectors for each variant: one VIP vector for the orthogonal components (Equations 8 and 13), one VIP vector for the predictive components (Equations 10 and 14), and one VIP vector for the global model (Equations 12 and 15).

$$VIP_{3,o} = \sqrt{\frac{K_o}{2} \times \left(\frac{\left[\sum_{a_o=1}^{A_o}\left(Wo_{a_o}^2 \times SSX_{comp,ao}\right)\right]}{SSX_{cum}} + \frac{\left[\sum_{a_o=1}^{A_o}\left(Wo_{a_o}^2 \times SSY_{comp,ao}\right)\right]}{SSY_{cum}} \right)} \tag{8}$$

$$K_o = \frac{K}{\frac{SSX_{cum,ao}}{SSX_{cum}} + \frac{SSY_{cum,ao}}{SSY_{cum}}} \tag{9}$$

$$VIP_{3,p} = \sqrt{\frac{K_p}{2} \times \left(\frac{\left[\sum_{a=1}^{A_p}\left(W_a^2 \times SSX_{comp,a}\right)\right]}{SSX_{cum}} + \frac{\left[\sum_{a=1}^{A_p}\left(W_a^2 \times SSY_{comp,a}\right)\right]}{SSY_{cum}} \right)} \tag{10}$$

$$K_p = \frac{K}{\frac{SSX_{cum,a}}{SSX_{cum}} + \frac{SSY_{cum,a}}{SSY_{cum}}} \tag{11}$$

$$VIP_{3,tot} = \sqrt{\frac{K}{2} \times \left(\begin{array}{c} \frac{\left[\sum_{a_o=1}^{A_o}\left(Wo_{a_o}^2 \times SSX_{comp,ao}\right)\right]}{SSX_{cum}} + \frac{\left[\sum_{a=1}^{A_p}\left(W_a^2 \times SSX_{comp,a}\right)\right]}{SSX_{cum}} + \\ \frac{\left[\sum_{a_o=1}^{A_o}\left(Wo_{a_o}^2 \times SSY_{comp,ao}\right)\right]}{SSY_{cum}} + \frac{\left[\sum_{a=1}^{A_p}\left(W_a^2 \times SSY_{comp,a}\right)\right]}{SSY_{cum}} \end{array} \right)} \tag{12}$$

$$VIP_{4,o} = \sqrt{\frac{K_o}{2} \times \left(\frac{\left[\sum_{a_o=1}^{A_o}\left(Po_{a_o}^2 \times SSX_{comp,\,ao}\right)\right]}{SSX_{cum}} + \frac{\left[\sum_{a_o=1}^{A_o}\left(Po_{a_o}^2 \times SSY_{comp,\,ao}\right)\right]}{SSY_{cum}} \right)} \tag{13}$$

$$VIP_{4,p} = \sqrt{\frac{K_p}{2} \times \left(\frac{\left[\sum_{a=1}^{A_p}\left(P_a^2 \times SSX_{comp,a}\right)\right]}{SSX_{cum}} + \frac{\left[\sum_{a=1}^{A_p}\left(P_a^2 \times SSY_{comp,a}\right)\right]}{SSY_{cum}} \right)} \tag{14}$$

$$VIP_{4,tot} = \sqrt{\frac{K}{2} \times \left(\begin{array}{c} \frac{\left[\sum_{a_o=1}^{A_o}\left(Po_{a_o}^2 \times SSX_{comp,ao}\right)\right]}{SSX_{cum}} + \frac{\left[\sum_{a=1}^{A_p}\left(P_a^2 \times SSX_{comp,a}\right)\right]}{SSX_{cum}} + \\ \frac{\left[\sum_{a_o=1}^{A_o}\left(Po_{a_o}^2 \times SSY_{comp,ao}\right)\right]}{SSY_{cum}} + \frac{\left[\sum_{a=1}^{A_p}\left(P_a^2 \times SSY_{comp,a}\right)\right]}{SSY_{cum}} \end{array} \right)} \tag{15}$$

Therefore, regardless of VIP variant (VIP_1–VIP_4), any variant chosen will yield three VIP vectors, one VIP for the predictive components (VIPpred), one VIP for the orthogonal components (VIPorth), and one VIP that is the total sum of the predictive and orthogonal parts (VIPtot). This allows an in-depth assessment of the variable influences for the three OPLS model compartments (the predictive, the orthogonal, and the global).

3. MATERIALS AND METHODS

The codes of the OPLS-VIP variants have been developed using MATLAB version R2013a (The MathWorks, Natick, MA, USA). The calculations have been tested and validated using SIMCA version 13.0 (MKS Umetrics AB, Umeå, Sweden) and MATLAB version R2013a (The MathWorks, Natick, MA, USA). The four OPLS-VIP

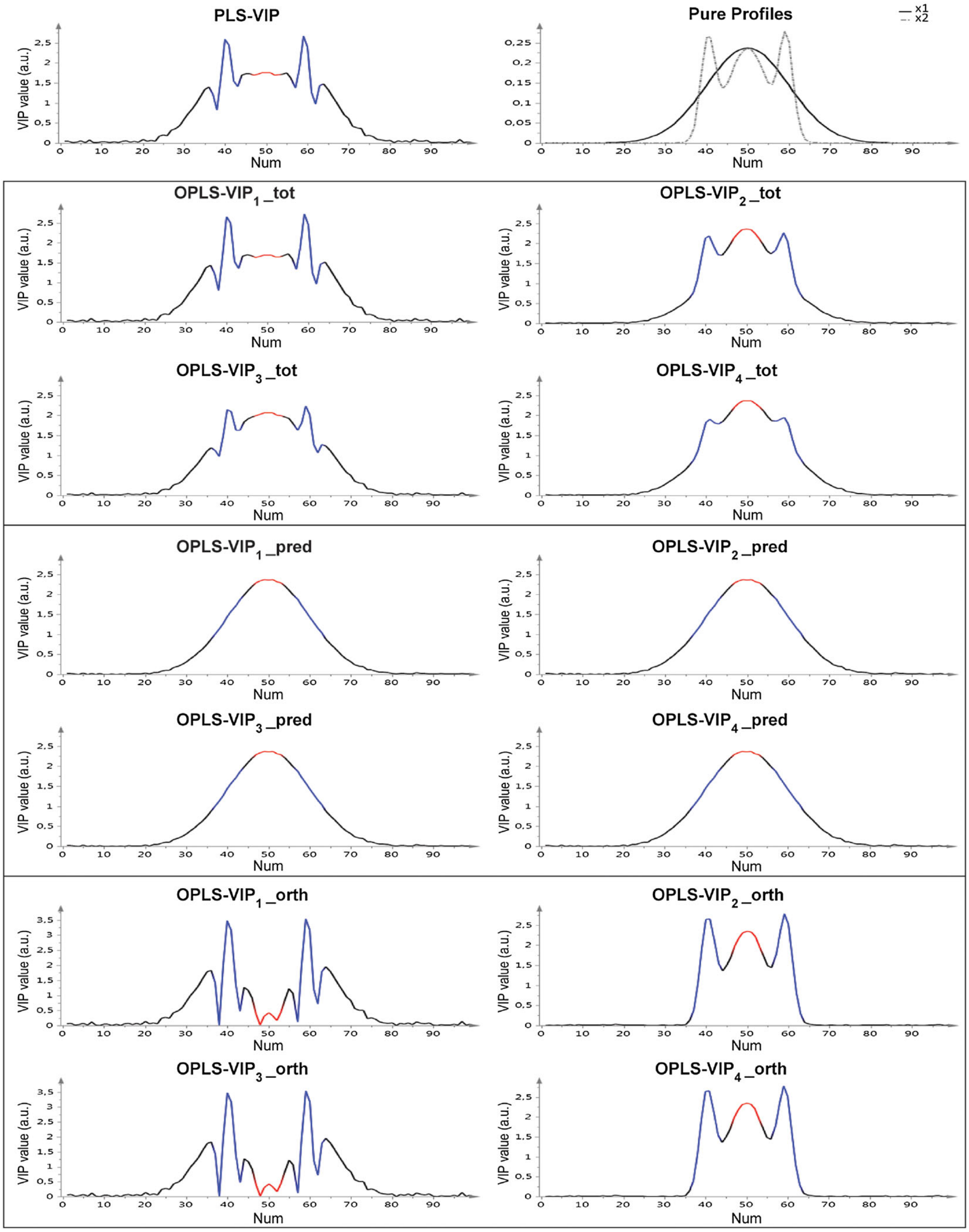

Figure 1. VIP results for two component single-y PLS model (top left figure) and 1 + 1 component single-y OPLS model using the PLS-VIP variant and the four OPLS-VIP variants, respectively, for the simulated data set. Results are grouped according to VIPtot, VIPpred, and VIPorth. Pure profiles are provided in the top right figure. Variables 47–53 are highlighted in red, and variables 37–43 and 57–63 are highlighted in blue, each representing the maximum peaks in the $\times_1$ and $\times_2$ profiles, respectively.

variants have been tested and compared using three data sets described in the following text.

3.1. Simulated data set

This simulated example has been previously described in the literature [22]. It comprises 100 variables and 70 observations. The simulated data set was constructed from two overlapping profiles in the variables, x_1 and x_2 (Figure 1), each normalized to length one. Their corresponding y_1 and y_2 vectors were orthogonal (i.e., 90° angle, zero correlation). The y_1-vector has equidistant values centered on zero and scaled to unit norm. The values in the y_2-vector were randomly generated, centered, and orthogonalized to y_1, and then scaled to unit norm. The X data matrix was calculated as the sum of both simulated components and a residual matrix E (contained about 1% of the total variance in X), as detailed in Equation 16.

$$X = y_1{}^*x_1^T + y_2{}^*x_2^T + E \tag{16}$$

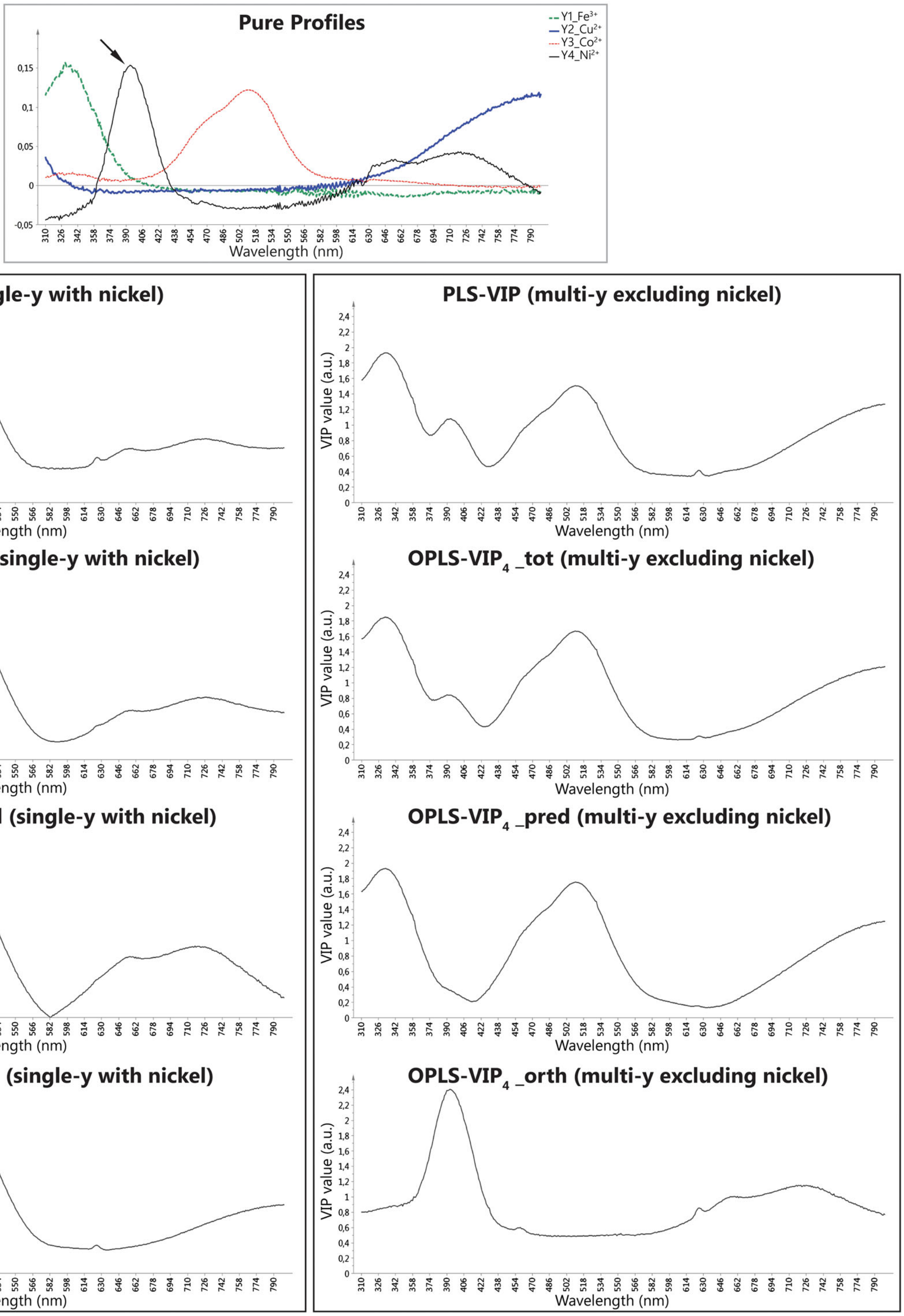

Figure 2. Comparison of PLS-VIP results with OPLS-VIP_4 results (OPLS-VIPtot, OPLS-VIPpred, and OPLS-VIPorth) using the loadings (p) and combining SSX and SSY for both predictive and orthogonal components. Pure spectral profiles of the four metal-ions complexes are provided in the top figure. The arrow points at the location of nickel peak (~390 nm).

3.2. Metal-ion data set

The metal-ion data set, which was used in the study of Trygg *et al*. [23,24], includes 52 mixtures of four different metal-ion complexes (Figure 2) that were mixed according to a design: $FeCl_3$ [0–0.25 mM], $CuSO_4$ [0–10 mM], $CoCl_2$ [0–50 mM], and Ni $(NO_3)^2$ [0–50 mM]. The design matrix does not have orthogonal columns, but it is still of full rank. The reference matrix consisted of 100 mM HCl. The mixtures were analyzed with a Shimadzu 3101PC UV–vis spectrophotometer in the wavelength region 310–800 nm, sampling at each wavelength to produce 491 variables. The data set was split into calibration and prediction sets with 26 observations in each; only the calibration set was required for the purpose of this paper. The preprocessing methods applied to the data were column mean-centering for X-variables and centering and scaling to unit variance for Y-variables. Both single-y (nickel metal-ion complex variable) and a multi-y (all metal-ion complex variables except nickel) models were evaluated. The nickel metal-ion complex variable was chosen for the single-y model because its pure profile has the largest overlap with the other profiles. This simulates a situation in which not all constituents in X are known in Y, that is, with orthogonal variation.

3.3. Wafers data set

The wafers data set comes from an oxide chemical vapor deposition process used in the manufacturing of a computer chip in the semiconductor industry [25]. The data set contains 3 months of process data collected across three similar process chambers. The three process chambers are supposedly identical in their performance, and it is of relevance to investigate whether this is the case. Chemical vapor deposition is a multistage process and can be treated like a batch process consisting of 12 phases. For each phase, many process variables were monitored, including gas flows, temperatures, pressures, and equipment settings such as angles and positions. In the current paper, we analyze the average value, per phase, for every process variable.

The wafers data set includes 2148 observations (wafers) processed by the three chambers. The wafers are distributed in this way: 675 wafers were processed by chamber A, 768 by chamber B, and 705 by chamber C. A total of 110 process variables were monitored in order to describe the production of the wafers. For proprietary reasons, the details of the variables cannot be disclosed.

An orthogonal projections to latent structures discriminant analysis (OPLS-DA) model was performed using the three chambers as the three classes for the modeling. All variables were column centered and scaled to unit variance.

4. RESULTS

For comparative purposes, PLS-VIP results are presented alongside the OPLS-VIP results.

4.1. Results for simulated data set

Two models were built using the single-y simulated data set to test the VIP variants: a two component single-y PLS model and a 1 + 1 single-y OPLS model. The parameters obtained from these models were used to calculate the VIP vectors, which afterwards were plotted in order to evaluate the four OPLS-VIP variants (Figure 1).

In order to facilitate the interpretation of the results, variables 47–53 were highlighted in red and variables 37–43 and 57–63 were highlighted in blue, each representing the maximum peaks in the x_1 and x_2 profiles, respectively. Comparing OPLS-VIP results and PLS-VIP results (Figure 1), it can be seen that the PLS-VIP and OPLS-VIP_1_tot plots (both of them based on *w* without [SSY,SSX] weighting) result in larger VIP values in the blue regions, which indicates that the conventional PLS-VIP and VIP_1 variant (based on *w* loadings) give similar profiles.

In OPLS-VIP_2_tot and OPLS-VIP_4 _tot, the maximum peak is found in the middle region (variables 47–53 highlighted in red). This is a result of the use of *p* loadings, making the middle region having large contributions from both predictive (OPLS-VIP_pred) and orthogonal (OPLS-VIP_orth) profiles.

4.2. Results for metal-ion data set

Five component models were calculated for both the single-y model (Ni^{2+}) and the multi-y (excluding Ni^{2+}) model. The same VIP evaluation was performed as for the simulated example except that we will now focus on OPLS-VIP_4 in comparison with PLS-VIP (Figure 2). The results using the other OPLS-VIP variants are found in Figures S2–S3.

For the single-y models, all VIP plots exhibit the peak of nickel (marked by an arrow in Figure 2) except the orthogonal OPLS-VIP_4_orth plot. The presence of the cobalt peak in the single-y PLS and OPLS plots is due to the fact that the cobalt and nickel concentrations are correlated, as observed by the correlation matrix where the correlation value between these two Y-variables was higher than the correlation value between the other Y-variables. The absolute correlation values in descending order are Co^{2+}–Ni^{2+} (0.384), Fe^{3+}–Cu^{2+} (0.124), Cu^{2+}–Ni^{2+} (0.103), Fe^{3+}–Ni^{2+} (0.093), Fe^{3+}–Co^{2+} (0.088), and Cu^{2+}–Co^{2+} (0.030).

Ni^{2+} was excluded in the multi-y models. As a result, it appears in the orthogonal OPLS-VIP_4_orth plot of the 3 + 2 multi-y OPLS model (and in consequence, also in the total OPLS-VIP_4_tot plot of the same model) at a wavelength region located around 390 nm (Figure 2).

4.3. Results for wafers data set

The resulting OPLS-DA model has two predictive components and three orthogonal components (2 + 3 OPLS-DA model). OPLS-VIP results were compared with PLS-VIP results obtained from a five-component PLS-DA model; see Figure 3 that again focuses on OPLS-VIP_4.

The PLS-VIP plot is shown at the top of Figure 3 and below the three different OPLS-VIP plots (VIPtot, VIPpred, and VIPorth, respectively). The variables that are considered more important in the PLSDA-VIP plot (marked in red) are the same as those found in the OPLSDA-VIPpred plot, but not in the OPLSDA-VIPorth plot. The variables that are more important for the orthogonal components (marked in blue) can only be elucidated using the new VIP for OPLS, as can be seen in Figure 3. Notice that the highlighting has been performed in two steps; firstly, red highlighting was performed on the PLSDA-VIP plot (top figure), and secondly, complementary blue highlighting was performed based on the OPLSDA-VIPorth plot (bottom figure). Consequently, all plots of Figure 3 show the important variables for the predictive components in red, and the important variables for the orthogonal components in blue. The plots resulting from the VIP vectors obtained using the other three variants (VIP_1–VIP_3) of OPLS-VIP can be found in Figure S4.

5. DISCUSSION

5.1. General considerations

We have outlined and exemplified four variants of OPLS-VIP, denoted VIP_1–VIP_4. For each VIP variant, we have described how to compute VIP vectors relating to predictive model components, orthogonal model components, and the total model. Thus, for each investigated dataset, this procedure has given rise to 12 different but OPLS-related VIP vectors (VIP_1–VIP_4 times three VIP types) plus the conventional original PLS-VIP, which was included for comparative purposes. In order to accomplish a summary and graphical overview of all 13 VIPs (12 OPLS-VIPs + 1 PLS-VIP), a hierarchical principal component analysis (hi-PCA) modeling can be

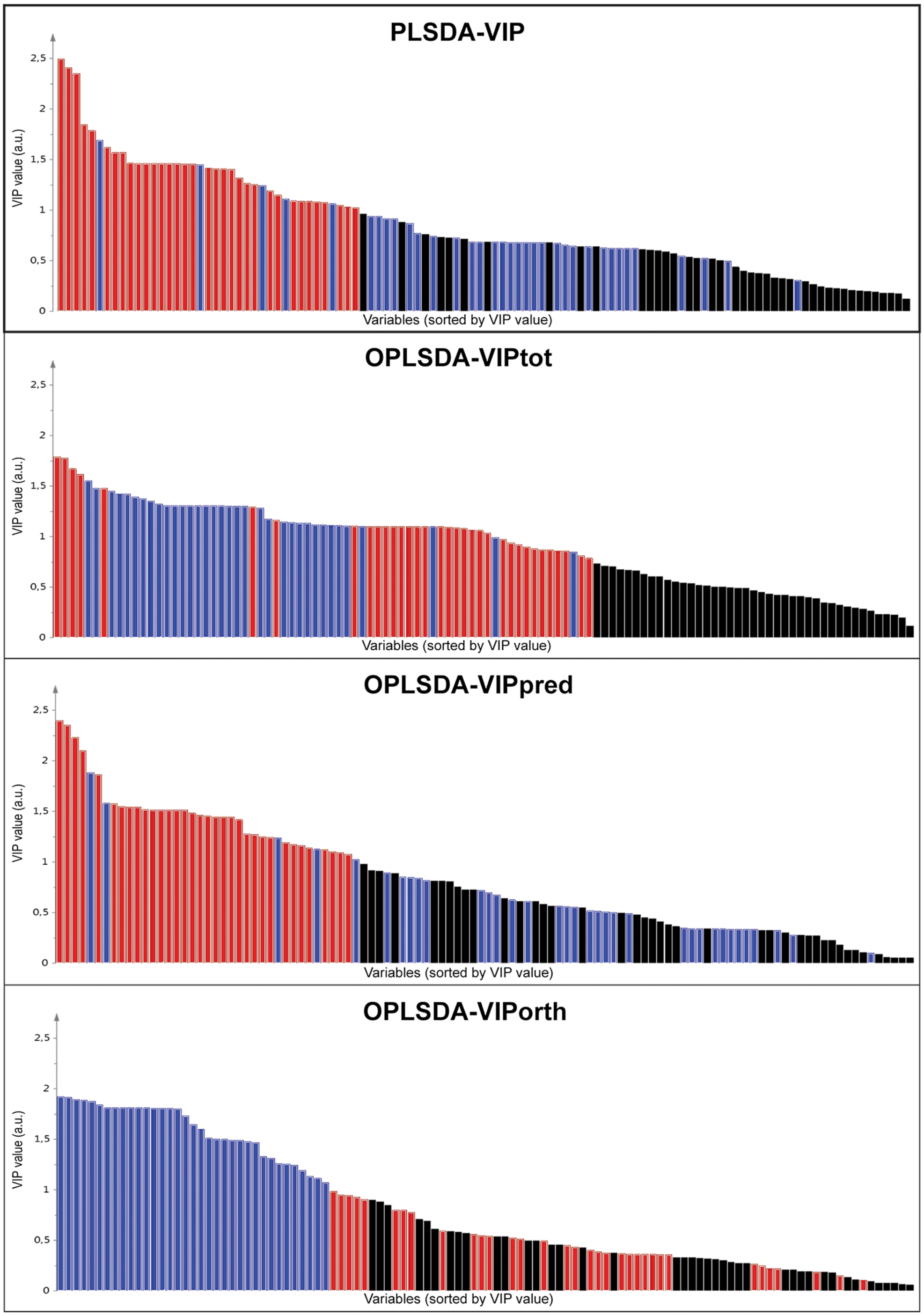

Figure 3. Wafers example. PLS-VIP plot of five component multi-y PLS-DA model (top figure) and OPLS-VIP_4 plot (VIPtot, VIPpred, and VIPorth) of a 2 + 3 component multi-y OPLS-DA model for wafers data set. Important variables for the predictive components are highlighted in red, whereas important variables for orthogonal components are highlighted in blue. Please note that the order of the variables of the plots of Figure 3 is not the same for all plots, the reason for this is that the variables have been sorted by descending VIP value.

performed [25,26]. The procedure for this is described in the Supporting Information, and the score scatter plot of the hierarchical principal component analysis (hi-PCA) model is displayed in Figure 4. The legend in Figure 4 indicates the type of VIP (OPLS-VIPpred, OPLS-VIPorth, and OPLS-VIPtot plus PLS-VIP). Several interesting observations can now be made. First of all, we can see that the orthogonal VIPs (green circle symbol in Figure 4), regardless of VIP_1–VIP_4 variant, stand out from the rest. This means the orthogonal VIPs encode new interpretative information, which is encouraging. Additionally, for the quartet of orthogonal VIPs, it can be deduced that the choice of loadings (normalized p or w) has more influence on the shapes of the VIP vectors compared with the choice of weighting principle (SSY and SSX separately as opposed to the combination [SSY,SSX]).

Furthermore, Figure 4 shows that the least spread occurs among the four predictive OPLS-VIPs (blue box symbol in Figure 4). This is according to expectation because they are predominantly affected by SSY and hence are conceptually analogous to the PLS-VIP. This

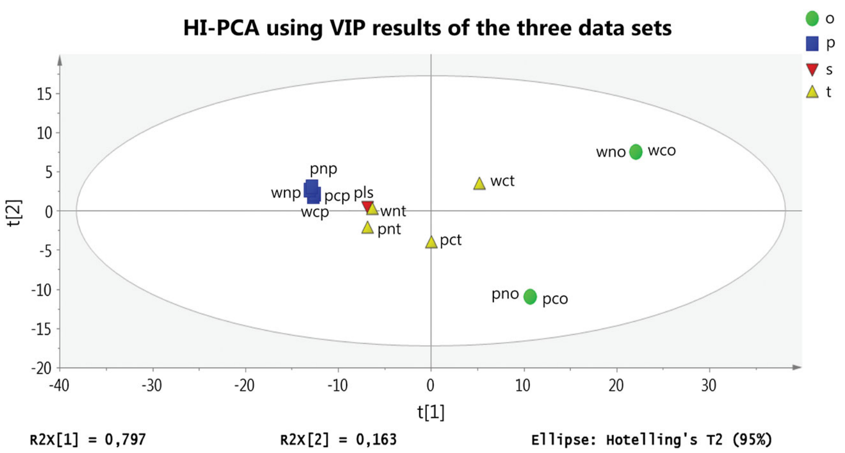

Figure 4. Score scatter plot of the hierarchical principal component analysis model built using the scores of five principal component analysis models that contain the VIP results using the three data sets. Labels of the points are coded according to the type of VIP, where the first letter (w/p) indicates the basis term used (loading weights or normalized loadings), the second letter (c/n) indicates if it is the conventional (c) weighting method using SSY and SSX separately or the new (n) weighting method using the combination [SSY,SSX], and the third letter (p/o/t/s) indicates if the VIP is related to the predictive (p) components, the orthogonal (o) components, the total (t) components, or PLS (s). The legend has been coded according to the third letter of the labels.

group of four predictive OPLS-VIPs is situated in the vicinity of the PLS-VIP point (red inverted triangle symbol in Figure 4). However, the two points that are sitting next to the PLS-VIP point arise from the total VIPs (yellow triangle symbol in Figure 4) of variants VIP$_3$ and VIP$_4$. This suggests that with the weighting of VIPs using the alloy of [SSY,SSX], we may mitigate against the extreme behavior of the orthogonal VIPs. Hence, if only an overview interpretation is sought for the total OPLS model, insights similar in nature to the PLS-VIP will be accomplished.

5.2. Discussion for simulated data set

The PLS-VIP and OPLS-VIP$_1$_tot plots in Figure 1 highlight VIP values in the marked blue regions resulting in misleading information, and both are based on *w* loadings in the calculations. In the blue regions, the x$_1$ profile does not encode such important variables, only the x$_2$ profile does. Interestingly, in the OPLS-VIP$_4$_tot plot, the relative importance of the variables is shown in a more realistic manner than in the plots of the other VIP variants; this can be appreciated in the relative sizes of blue and red regions in both plots. So, VIP$_4$ (which is calculated using *p* and [SSY,SSX] combination) is the variant that leads to more realistic and reliable results.

5.3. Discussion for metal-ion data set

Comparing the plots of VIPtot and VIPorth of the multi-y OPLS model using all Y-variables (Figure S1) and the multi-y OPLS model excluding nickel (Figure S2), we realize that the pure spectral profile of the nickel variable has a large overlap with the other spectral profiles.

Taking a global view of the results for the metal-ion data set and comparing with the pure profiles, variant OPLS-VIP$_4$ (based on *p* and [SSY,SSX] combination) seems to give more informative results, especially when orthogonal variance is present in the model. Results obtained for predictive, orthogonal and total VIP$_4$ show in a clear and realistic manner which variables are more relevant in each case. Thus, the novel VIP for OPLS allows us to see the results separately for predictive and orthogonal components, which will aid in the total understanding of the properties of the data set.

5.4. Discussion for wafers data set

In the wafers example (Figure 3), from the comparison between the OPLSDA-VIP$_4$ plots and the PLSDA-VIP plot, it can be deduced that the PLS-VIP provides good results for the predictive components but not for the orthogonal components; this is due to the fact that the weighting of the conventional PLS-VIP formula is only sensible for the predictive components but not for the orthogonal components. On the other hand, the new OPLS-VIP variant, which uses normalized loadings (*p*) and the combination [SSY,SSX] for both predictive and orthogonal components, uncovers important variables for the orthogonal components, which escaped undetected by the conventional PLS-VIP.

The three process chambers (denoted A, B, and C) are supposedly identical in their processing of the wafers. However, contrary to expectation, there is a strong and systematic difference between the three chambers [23]. A detailed analysis of the OPLS-DA model is beyond the scope of the current paper but is given elsewhere [23]. Close inspection of product quality measurement (oxide layer thickness) show that the wafers processed with chamber B invariably obtained slightly thinner oxide layers.

Interpreting the OPLS-VIP$_4$pred vector will indicate which process variables correlate strongly to the analysis question, that is, to separate the chambers. These process variables are colored in red in Figure 3. For these variables, there are systematic differences between the three process chambers. Some examples are process variables reflecting ozone concentration measurements and equipment settings parameters.

Moreover, interpreting the OPLS-VIP$_4$orth vector will identify process variables in which there is systematic variation among wafers, or subsets of wafers, but which is not connected to the analysis question (to separate the chambers). Such process variables are colored blue in Figure 3, and for these, there is no systematic difference between the chambers. Some examples are process variables reflecting various temperature and pressure measurements.

6. CONCLUDING REMARKS

According to the results presented in this paper, the OPLS-VIP calculated using normalized loadings (*p*) and [SSY,SSX] combination (denoted VIP_4) improves the OPLS model interpretability. In the same way that OPLS explains the variation of the predictive and the orthogonal components, the OPLS-VIP_4 presented here can point at the variables that are more important for both predictive and orthogonal components. Additionally, this innovative OPLS-VIP gives the results in two clear ways: three VIP vectors (predictive VIP vector, orthogonal VIP vector, and total VIP vector) and three intuitive VIP plots (VIPpred, VIPorth, and VIPtot). Thus, this new VIP for OPLS allows us to see the results separately for predictive and orthogonal components, what is a clear advantage for the model interpretation. We stress that VIP for the orthogonal components stands out (Figure 4) representing new interpretative information. In summary, the OPLS-VIP_4 variant lays the ground for a model interpretation that is well in line with the structure of the underlying data. Accordingly, this alternative is the preferred one.

As demonstrated by means of the examples, the VIP vectors (VIPpred, VIPorth, and VIPtot) predominantly are plotted as line plots or bar plots; the choice of the plot type depends on the nature of the data. In the second data set, there exists a physical ordering structure among the variables, that is, the wavelengths, which implies that the line plot representation is the logical one. In the third data set, however, no such obvious ordering structure exists among the different process variables. In such a case, sorting the values of a particular VIP vector according to decreasing numerical value ("size-sorting") is the common practice. One can also envision extensions to these basic plot types, for instance, coloring according to another model parameter. Such ideas are in the pipeline and will be exploited in future works.

Since its inception [3], the classical VIP parameter for PLS has been used as a compact representation for model interpretation across all Y-variables. However, as pointed out by one of the anonymous reviewers and as shown in [11], the classical PLS-VIP may well be re-expressed to cover only one Y-variable at a time. The same holds true for the OPLS-VIP codes (VIPpred, VIPorth, and VIPtot) outlined in this article; instead of applying to all Y-variables, they can be re-formulated to cover just one Y-variable at a time. In fact, this sub-division principle can be extended into the multi-block situation, in which three or more datablocks are used in a data integration and comparison objective. When handling three or more datablocks, one global OPLS-VIP of any type (VIPpred, VIPorth, and VIPtot) will not provide enough detail to understand the complete pattern of information overlap between the various datablocks. Let us consider the three-block situation with datablocks X, Y, and Z. Clearly, a global X-Y-Z related VIP of any type (VIPpred, VIPorth, and VIPtot) will need to be supplemented by derivatives thereof, for example, local X-Y, X-Z, and Y-Z related VIPs. Hence, at least conceptually, there is some resemblance between in the two-block situation dividing a global VIP into individual Y-variables and in the multi-block situation dividing a global VIP into local two-block counterparts. More studies into this are planned, and we hope to report our results in the near future.

Acknowledgements

The authors would like to acknowledge the financial support from the Swedish Research Council (JT) grant no. 2011–604, the MKS Umetrics AB (BG-P), and the Industrial Doctoral School (IDS), Umeå University, Sweden. In addition, the anonymous reviewers are thanked for their valuable comments to this paper.

REFERENCES

1. Wold S, Martens H, Wold H. The multivariate calibration-problem in chemistry solved by the PLS method. *Lect Notes Math* 1983; **973**: 286–293.
2. Geladi P. Kowalski BR. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* ,1986; **185**: 1–17.
3. Wold S, Johansson E, Cocchi M. PLS—partial least-squares projections to latent structures. *3D QSAR in Drug Design*, Theory Methods and Applications, Kubinyi H (eds.). ESCOM Science Publishers: Leiden, 1993; 523–550.
4. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Sys* 2001; **58**(2): 109–130.
5. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemometr* 2002; **16**(3): 119–128.
6. Svensson O, Kourti T, MacGregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. *J Chemometr* 2002; **16**(4): 176–188.
7. Ergon R. PLS post-processing by similarity transformation (PLS plus ST): a simple alternative to OPLS. *J Chemometr* 2005; **19**(1): 1–4.
8. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M, Kvalheim OM. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal Chem* 2009; **81** (7): 2581–2590.
9. Pinto RC, Trygg J, Gottfries J. Advantages of orthogonal inspection in chemometrics. *J Chemometr* 2012; **26**(6): 231–235.
10. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu T-M, Goodsaid FM, Pusztai L, Shaughnessy JD Jr, Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, Chen M, Cheng J, Cheng J, Chou J, Davison TS, Delorenzi M, Deng Y, Devanarayan V, Dix DJ, Dopazo J, Dorff KC, Elloumi F, Fan J, Fan S, Fan X, Fang H, Gonzaludo N, Hess KR, Hong H, Huan J, Irizarry RA, Judson R, Juraeva D, Lababidi S, Lambert CG, Li L, Li Y, Li Z, Lin SM, Liu G, Lobenhofer EK, Luo J, Luo W, McCall MN, Nikolsky Y, Pennello GA, Perkins RG, Philip R, Popovici V, Price ND, Qian F, Scherer A, Shi T, Shi W, Sung J, Thierry-Mieg D, Thierry-Mieg J, Thodima V, Trygg J, Vishnuvajjala L, Wang SJ, Wu J, Wu Y, Xie Q, Yousef WA, Zhang L, Zhang X, Zhong S, Zhou Y, Zhu S, Arasappan D, Bao W, Lucas AB, Berthold F, Brennan RJ, Buness A, Catalano GJ, Chang C, Chen R, Cheng Y, Cui J, Czika W, Demichelis F, Deng X, Dosymbekov D, Eils R, Feng Y, Fostel J, Fulmer-Smentek S, Fuscoe JC, Gatto L, Ge W, Goldstein DR, Guo L, Halbert DN, Han J, Harris SC, Hatzis C, Herman D, Huang J, Jensen RV, Jiang R, Johnson CD, Jurman G, Kahlert Y, Khuder SA, Kohl M, Li J, Li L, Li M, Li Q-Z ,Li S, Li Z, Liu J, Liu Y, Liu Z, Meng35 L, Madera M, Martinez-Murillo F, Medina I, Meehan J, Miclaus K, Moffitt RA, Montaner D, Mukherjee P, Mulligan GJ, Neville P, Nikolskaya T, Ning B, Page P Grier, Parker J, Parry RM, Peng X, Peterson RL, Phan JH, Quanz B, Ren Y, Riccadonna S, Roter AH, Samuelson FW, Schumacher MM, Shambaugh JD, Shi Q, Shippy R, Si S, Smalter A, Sotiriou C, Soukup M, Staedtler F, Steiner G, Stokes TH, Sun Q, Tan P-Y, Tang R, Tezak Z, Thorn B, Tsyganova M, Turpaz Y, Vega SC, Visintainer R, Frese JV, Wang C, Wang E, Wang J, Wang W, Westermann F, Willey JC, Woods M, Wu S, Xiao N, Xu J, Xu L, Yang L, Zeng X, Zhang J, Zhang L, Zhang M, Zhao C, Puri RK, Scherf U, Tong W, Wolfinger RD. The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotech* 2010; **28**(8): 827–838.
11. Favilla S, Durante C, Li Vigni M. Cocchi M Assessing feature relevance in NPLS models by VIP. *Chemom Intell Lab Sys* 2013; **129** (15): 76–86.
12. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J Chemometr* 2010; **24**(11–12): 728–737.
13. Teofilo RF, Martins JPA, Ferreira MMC. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J Chemometr* 2009; **23**(1–2): 32–48.
14. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Sys* 2005; **78** (1–2): 103–112.

15. Mohajeri A, Hemmateenejad B, Mehdipour A, Miri R. Modeling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analyzed by GA-PLS and PC-GA-PLS. *J Mol Graph Model* 2008; **26**(7): 1057–1065.
16. Sun HM. Prediction of chemical carcinogenicity from molecular structure. *J Chem Inf Comput Sci* 2004; **44**(4): 1506–1514.
17. Han PP, Yuan YJ. Lipidomic analysis reveals activation of phospholipid signaling in mechanotransduction of Taxus cuspidata cells in response to shear stress. *FASEB J* 2009; **23**(2): 623–630.
18. Rossel RAV, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 2010; **158**(1–2): 46–54.
19. Koukoulitsa C, Tsantili-Kakoulidou A, Mavromoustakos T, Chinou I. PLS analysis for antibacterial activity of natural coumarins using VolSurf descriptors. *QSAR Comb Sci* 2009; **28**(8): 785–789.
20. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. Multi- and megavariate data analysis (part 2). Second ed. Vol. Advanced Applications and Method Extensions. 2006: p. 266.
21. Kvalheim OM, Rajalahti T, Arneberg R. X-tended target projection (XTP)-comparison with orthogonal partial least squares (OPLS) and PLS post-processing by similarity transformation (PLS plus ST). *J Chemometr* 2009; **23**(1–2): 49–55.
22. Stenlund H, Johansson E, Gottfries J, Trygg J. Unlocking interpretation in near infrared multivariate calibrations by orthogonal partial least squares. *Anal Chem* 2009; **81**(1): 203–209.
23. Trygg J. Prediction and spectral profile estimation in multivariate calibration. *J Chemometr* 2004; **18**(3–4): 166–172.
24. Trygg J. O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J Chemometr* 2002; **16**(6): 283–293.
25. Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C. Multi- and megavariate data analysis. Third revised ed. Vol. Basic Principles and Applications. 2013: p. 224–229, 355–371.
26. Wold S, Kettaneh N, Tjessem K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J Chemometr* 1996; **10**(5–6): 463–482.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.