Lee and Lee *BioData Mining* (2021) 14:7
https://doi.org/10.1186/s13040-021-00240-3

BioData Mining

METHODOLOGY Open Access

Feature selection using distributions of orthogonal PLS regression vectors in spectral data

Geonseok Lee and Kichun Lee*

*Correspondence:
skylee@hanyang.ac.kr
Industrial Engineering, Hanyang University, Seoul, Korea

Abstract

Feature selection, which is important for successful analysis of chemometric data, aims to produce parsimonious and predictive models. Partial least squares (PLS) regression is one of the main methods in chemometrics for analyzing multivariate data with input *X* and response *Y* by modeling the covariance structure in the *X* and *Y* spaces. Recently, orthogonal projections to latent structures (OPLS) has been widely used in processing multivariate data because OPLS improves the interpretability of PLS models by removing systematic variation in the *X* space not correlated to *Y*. The purpose of this paper is to present a feature selection method of multivariate data through orthogonal PLS regression (OPLSR), which combines orthogonal signal correction with PLS. The presented method generates empirical distributions of features effects upon *Y* in OPLSR vectors via permutation tests and examines the significance of the effects of the input features on *Y*. We show the performance of the proposed method using a simulation study in which a three-layer network structure exists in compared with the false discovery rate method. To demonstrate this method, we apply it to both real-life NIR spectra data and mass spectrometry data.

Keywords: Feature selection, PLS, Orthogonal signal correction, Regression vector, Permutation test

Introduction

Feature selection is a technique to select a subset of variables which are useful in predicting target responses. From a machine learning viewpoint, irrelevant features in a prediction model deteriorate its generalization ability, and it is critical to remove such redundant features to keep the model from being misled by inappropriate information. The task of feature selection, one of the central tasks in machine learning, helps to reduce overfitting by eliminating redundant features. The prevalence of high-dimensional data becomes a challenge for researchers to perform feature selection [1]. In biological fields,

particulary, high dimensionality often arises, and irrelevant and redundant features make up a high proportion of the total data [2].

Indeed, contemporary analytic methods such as near-infrared (NIR), proton nuclear magnetic resonance (^{1}H NMR) spectroscopy, liquid chromatography-mass spectrometry (LC-MS), and gas chromatography-mass spectrometry (GC-MS) provide high-dimensional data sets in which the number of features is usually larger than the number of observations. Those spectral data sets, denoted by the input variables (feature) X, are an effective alternative to using classical chemical analyses in screening experiments [3]. They can reveal the underlying patterns associated with health characteristics, the so-called phenotypes. These include pathological characteristics, denoted by the response variables Y, and thus can be of substantial value in biomedical research [4]. To this end, reliable identification of features associated with response characteristics is important.

Uncovering hidden patterns associated with the response variables in spectral data sets is not trivial. One of the major problems in addressing this issue is how to deal with the existence of spectral collinearity. Spectral collinearity, originating from linear dependence among the input variables X, poses ill-conditioned linear equation systems. Thus, standard regression methods cannot be applied in this context, so different strategies need to be adopted. The methods usually employed for solving such problems are artificial neural networks, k-nearest neighbors, principle component analysis (PCA), and partial least squares (or projections to latent structures, PLS) among others. A detailed discussion of multivariate statistical methods in spectral data sets can be found in Holmes and Antti [5] and Lindon et al. [6].

Among these, unsupervised PCA and supervised PLS together with regression have been widely applied. These methods are useful for reducing the complexity in the feature space, the main idea is to find a low dimensional representation while retaining as much of the variation as possible. Compared to other methods, they are not only easy to interpret but also effective in explaining the interrelationships among spectral data by examining the variance levels of spectral data while being less computationally demanding. Basically, they produce a few transformed scores (positions for new directions) for the original input variables X to reduce the complexity of such data and retain most of the variational information.

For the task of identifying significant variables, however, produced scores by PCA often undergo lack of discrimination power since it only focuses on new directions or loadings, accounting for the maximum variation of X, and then projects the original input variables onto the new directions [7]. On the other hand, PLS is a type of regression model used to find the relationships between the response variables and input variables based on the assumption that they are generated by a common set of underlying factors [8]. That is to say, it finds the directions in the space of X that explains the maximum of variation of the space Y. By reducing the collinearity between input variables and increasing covariance between input and response variables at the same time, feature selection by PLS can result in a more parsimonious model without losing its predictive ability.

The task of identifying significant features via PLS faces a few challenges. The selection of original spectral variables from those transformed scores is nontrivial because they are represented as linear combinations of a large number of the original input variables. PLS loadings express the weights for these linear combinations and generate PLS regression vectors with estimated regression coefficients. The uncertainty in the estimates of

PLS loadings and scores in conjunction with PLS regression vectors complicates this task. Since error in a few original input variables are propagated in all transformed scores by way of loadings, the uncertainty in PLS regression vectors increases accordingly [9, 10]. Additionally, the absence of closed-form distributions of PLS loadings also makes the task challenging. To cope with these challenges in PLS-based variable selection, several statistical approaches have been suggested. Heise applied cross-validation with a strategy of expanding neighboring variables [11]. Høskuldsson employed the stepwise regression approach based on the goodness of fit criterion [12]. Faber proposed a resampling technique [13], focusing on variables interval selection. These approaches, though known to be unbiased [14], inevitably suffer from uncertainty under a substantial number of features and often lead to the selection of unnecessary features. The presence of unnecessary features, causing overfitting, is a critical issue especially when the number of observations is less than that of features. It would be a benefit to apply a resampling method, able to regulate the selection of unnecessary features, to the identification of significant features in PLS.

This paper introduces a useful combined approach of applying orthogonal signal correction (OSC) and permutation tests to PLS for the purpose of feature selection. OSC, introduced by Wold et al. [15], removes from input variables only the part unrelated to the response variable. Extensions of OSC aiming to improve model prediction and interpretability were also presented [16, 17], sharing the same spirit in that X variation unrelated to the response is filtered out as a preprocessing step. Thus in the combined approach, first, OSC as a preprocessing step of PLS corrects X by removing systematic variation in the X space not correlated to Y. Second, we relate the orthogonal-signal corrected X to the PLS regression models, which we call OPLS models, and obtain OPLS-induced regression vectors. The regression vectors in OPLS models describe each variable's contribution to the response Y while reflecting variables collectively on new directions. Lastly, by employing a permutation testing procedure, in which permutations of the input observations for a collection of randomly chosen features occurs, we test the significance of each coefficient in the regression vectors in a reasonably fast manner. The permutation test procedure yields empirical null distributions of weights for each individual variable's effect on the response Y. The contribution of this paper is the integration of permutation test into feature selection of OPLS models combining the concepts of OSC and PLS. It then introduces the use and testing of OPLSR vectors for the purpose of feature selection, particularly in the domains of spectral data.

This paper is organized as follows. "Proposed method" section describes the combined method of orthogonal signal correction and PLS, manifesting the adopted permutation test procedure. It also includes a simulation study in which simulated data sets with a three-level network structure are used and the performance of the proposed approach is demonstrated. The proposed approach is compared with two methods: one is the Lasso, penalized linear model, and the other is the false discovery rate (FDR) method which is based on the variables' individual effects on the response Y and is widely used in spectral data analysis [18, 19]. "Experiments" section demonstrates the approach with a real-life NIR data set. Finally, "Conclusions" section concludes the paper.

Proposed method

We describe orthogonal signal correction and PLS to obtain OPLS regression vectors.

Then we introduce a permutation-based test procedure to test the significance of each variable's effect in OPLS regression vectors.

Orthogonal PLS regression vector

In the analysis of chemometric data using a multivariate regression model $Y = XB + E$, it is common that the first component accounting for the highest percentage of the input X variation constitutes only a low percentage of the response Y variation. The input X and response Y are assumed to be mean-centered and properly normalized. In PCA, a principal component vector (or score vector) $\mathbf{t}$ is calculated by a linear combination of an input observation $\mathbf{x}_j, j = 1, \ldots, n$ (row of X) and a loading vector $\mathbf{p}$. Score vector $\mathbf{t}$ can be regarded as a transformed variable by the corresponding loading vector. Loading vector $\mathbf{p}$ is numerically found as an eigenvector of the sample covariance matrix $X^T X/(n-1)$. PLS regression, however, focuses on the covariance structure between X and Y. In PLS regression, score vectors and loading vectors for X and Y are jointly and numerically obtained to maximize the covariance between X score vectors and Y score vectors [20].

The goal of orthogonal signal correction (OSC) is to remove one or more directions in X ($n \times p$) that are unrelated, or more precisely orthogonal, to Y and to account for the largest variation in X as well. OSC often serves as a pre-processing step to improve the multivariate regression model [21]. Particularly, the PLS regression coefficients after OSC have stronger interpretability and clarity. In this paper, we chose direct orthogonal signal correction since it not only bears close relations with other OSC methods and works quite well with empirical data, but also uses only least squares methods without iterative computation [22]; this means it can be analytically connected with subsequent PLS. We summarize the basic steps as follows and link them to PLS to obtain orthogonal-signal-corrected PLS regression (OPLSR) vectors that will be used for feature selection.

The first step is to take the projection of Y onto X, $\hat{Y} = \mathbf{P}_X Y$, where $\mathbf{P}_X$ represents a projection matrix to the column space of X, denoted by $C(X)$. Furthermore, we write Y as follows:

$$Y = \mathbf{P}_X Y + (Y - \mathbf{P}_X Y) := \mathbf{P}_X Y + \mathbf{A}_X Y = \hat{Y} + \mathbf{A}_X Y,$$

where $\mathbf{A}_X = I - \mathbf{P}_X$ represents a projection matrix to the space orthogonal to the column space of X. Accordingly, $\mathbf{A}_X Y$ is orthogonal to $C(X)$: for $\mathbf{v} \in C(X)$,

$$\mathbf{v}^\top \mathbf{A}_X Y = 0. \tag{1}$$

For underdetermined systems, $p > n$, as is common in spectral data, $\hat{Y}$ equals Y.

The second step is to decompose X into two orthogonal parts, one part that has the same range $\hat{Y}$ and another that is orthogonal to it:

$$X = \mathbf{P}_{\hat{Y}} X + (X - \mathbf{P}_{\hat{Y}} X) = \mathbf{P}_{\hat{Y}} X + \mathbf{A}_{\hat{Y}} X.$$

Thus, the space spanned by the columns of $\mathbf{A}_{\hat{Y}} X$, which is essentially a subspace in X and orthogonal to $\hat{Y}$, is a target of removal from X in the OSC procedure. The space $\mathbf{A}_{\hat{Y}} X$ is also orthogonal to Y: for $\mathbf{v} \in C(\mathbf{A}_{\hat{Y}} X)$,

$$Y^\top \mathbf{v} = (\hat{Y} + \mathbf{A}_X Y)^\top \mathbf{v} = \hat{Y}^\top \mathbf{v} + (\mathbf{A}_X Y)^\top \mathbf{v} = 0, \tag{2}$$

since $\hat{Y}$ is orthogonal to $\mathbf{A}_{\hat{Y}} X$ by definition and $\mathbf{v}$, also in $C(X)$, is orthogonal to $\mathbf{A}_X Y$ by (1).

The third step is to find the first principal score vector $\mathbf{t}_{OSC}$ and the associated loading vector $\mathbf{p}_{OSC}$ from $\mathbf{A}_{\hat{Y}}X$, so that $\mathbf{t}_{OSC}\mathbf{p}_{OSC}^{\top}$ approximates to $\mathbf{A}_{\hat{Y}}X$. Let X^{ORTH} be the approximation of $\mathbf{A}_{\hat{Y}}X$ by $\mathbf{t}_{OSC}\mathbf{p}_{OSC}^{\top}$: $X^{ORTH} = \mathbf{t}_{OSC}\mathbf{p}_{OSC}^{\top}$. The score vector $\mathbf{t}_{OSC}$, a direction or an OSC component, accounts for the maximum variance of $\mathbf{A}_{\hat{Y}}X$. Moreover, the score vector $\mathbf{t}_{OSC} \in \mathbf{A}_{\hat{Y}}X$ is orthogonal to $\hat{Y}$ by the definition of $\mathbf{A}_{\hat{Y}}X$ and also orthogonal to Y by (2):

$$\mathbf{t}_{OSC}^{\top}Y = 0. \tag{3}$$

Being in $C(X)$, the score vector $\mathbf{t}_{OSC}$ is expressed as a linear combination of the columns of X with a weight vector $\mathbf{r}_{OSC}$:

$$\begin{aligned} &\mathbf{t}_{OSC} = X\mathbf{r}_{OSC}, \text{ or} \\ &\mathbf{r}_{OSC} = X^{\dagger}\mathbf{t}_{OSC}, \end{aligned} \tag{4}$$

where $X^{\dagger}$ is the Moore-Penrose pseudoinverse of X. Overall, the orthogonal-signal corrected X preserving predictive components, denoted by X^{OSC}, writes:

$$X^{OSC} := X - X^{ORTH} = X - \mathbf{t}_{OSC}\mathbf{p}_{OSC}^{\top} = X - X\mathbf{r}_{OSC}\mathbf{p}_{OSC}^{\top}. \tag{5}$$

Here, weight vector $\mathbf{r}_{OSC}$ consists of weights of individual variables contributions to the construction of a space orthogonal to Y, which will be used to relate the input matrix X to the OPLSR vectors in the last step. Practically, it is necessary to limit the number of OSC components. To avoid over-fitting, which results in a loss of model generality, the number of OSC components should not be too high. Mostly, one or two OSC components are sufficient. The first OSC component often describes a base-line correction, and the second can serve as the correction of multiplicative effects [22]. In this study, we chose one OSC component in consideration of the intuitive interpretation and the sufficient amount of its variability in X in practice.

Now that X^{OSC} is found, we apply PLS to X^{OSC} and Y. The reason PLS is applied to X^{OSC} and not directly to X is that the covariance between X and Y equals that between X^{OSC} and Y:

$$X^{\top}Y = \left(X^{OSC} + \mathbf{t}_{OSC}\mathbf{p}_{OSC}^{\top}\right)^{\top} Y = \left(X^{OSC}\right)^{\top} Y + \mathbf{p}_{OSC}\mathbf{t}_{OSC}^{\top}Y = \left(X^{OSC}\right)^{\top} Y,$$

because $\mathbf{t}_{OSC}$ is orthogonal to Y by (3). Thus, the PLS finds a few score vectors of X^{OSC} that maximize the covariance between X^{OSC} score vectors and Y score vectors. The scores of X^{OSC}, $\mathbf{t} = [\,\mathbf{t}_1\ \mathbf{t}_2\ \cdots\ \mathbf{t}_A\,]$ (of size $n \times A$) corresponding to A predictive components, are not only orthogonal, but also obtainable by computing its associated weight vectors $\mathbf{r} = [\,\mathbf{r}_1\ \mathbf{r}_2\ \cdots\ \mathbf{r}_A\,]$ from the singular value decomposition of the covariance matrix $(X^{OSC})^{\top}Y$:

$$\begin{aligned} &\mathbf{t}_j = X^{OSC}\mathbf{r}_j, \quad j = 1, \cdots, A, \text{ or} \\ &\mathbf{t} = X^{OSC}\mathbf{r}. \end{aligned} \tag{6}$$

Accordingly, the corresponding loading of X^{OSC}, denoted by P, is given by $P = \left(X^{OSC}\right)^{\top}\mathbf{t}$, meaning new weights of X^{OSC} are directly evaluated by $\mathbf{t}$. Similarly, the loading of Y, denoted by Q, is computed by $Q = Y^{\top}\mathbf{t}$ and represents new weights of Y that lead to maximization of the covariance matrix using the A predictive components commonly applied to X^{OSC} and Y [23].

Overall, X^{OSC} and Y are decomposed into the following:

$$X^{OSC} = \mathbf{t}P^\top + E,$$
$$Y = \mathbf{u}Q^\top + F,$$

where $\mathbf{u}$ is the score matrix of Y given by $\mathbf{u} = YQ$; and E and F are residual matrices.

By regressing Y on the score $\mathbf{t}$, which is a compact representation of X^{OSC} based on the A predictive components, from the regression model $Y \sim \mathbf{t}\beta$, the least-squares estimate of the coefficient β becomes $\hat{\beta} = (\mathbf{t}^\top\mathbf{t})^{-1}\mathbf{t}^\top Y = \mathbf{t}^\top Y$ due to the orthogonality of $\mathbf{t}$. Then the prediction model $\hat{Y}$ is

$$\hat{Y} = \mathbf{t}\hat{\beta} = \mathbf{t}\mathbf{t}^\top Y = X^{OSC}\mathbf{r}\mathbf{t}^\top Y = X^{OSC}\mathbf{r}Q^\top := X^{OSC}\hat{\beta}_{OPLSa}, \tag{7}$$

where $\hat{\beta}_{OPLSa} = \mathbf{r}Q^\top$ is the regression vector using the orthogonal signal corrected X, X^{OSC}. Furthermore, the prediction model $\hat{Y}$ can be rewritten as follows:

$$\hat{Y} = X^{OSC}\hat{\beta}_{OPLSa} = \left(X - X\mathbf{r}_{OSC}\mathbf{p}_{OSC}^\top\right)\hat{\beta}_{OPLSa} := X\hat{\beta}_{OPLSb}, \tag{8}$$

where $\hat{\beta}_{OPLSb} = \left(I - \mathbf{r}_{OSC}\mathbf{p}_{OSC}^\top\right)\hat{\beta}_{OPLSa}$ is the regression vector based on X. Both $\hat{\beta}_{OPLSa}$ and $\hat{\beta}_{OPLSb}$ consist of weights of individual variables i, $i = 1, \cdots, p$, contributing to the change of Y. In other words, a large absolute value of $\hat{\beta}_{OPLSa,i}$ or $\hat{\beta}_{OPLSb,i}$ indicates that the ith variable (x_i) contributes substantially to the increase or decrease of Y depending on the sign. We notice that the use of $\hat{\beta}_{OPLSb,i}$ in contrast to $\hat{\beta}_{OPLSa,i}$ brings comprehensive examination of individual variable effects because it includes the presence of all X directions, both orthogonal and predictive. Thus we propose $\hat{\beta}_{OPLSb}$ as the final regression coefficient vector, regarding $\hat{\beta}_{OPLSa}$ as a side source for comparison. This comparison will be demonstrated in a simulation study and a real-life data example that follow. We will obtain the distributions of $\hat{\beta}_{OPLSa}$ and $\hat{\beta}_{OPLSb}$ to systemically select variables that significantly contribute to the change of Y.

Permutation tests

Since the OPLS regression vectors, $\hat{\beta}_{OPLSa}$ and $\hat{\beta}_{OPLSb}$ are computed, the aim here is to test the significance of individual variables effects on the vectors. Let $\hat{\beta}_{OPLS}$ denote the two regression vectors generally. Permutation tests, a computer-based re-sampling method for achieving accuracy measures of statistical estimates from an approximating distribution [24], are widely used in the computation of variable importance and confidence intervals in random forests [25, 26].

The advantage of permutation tests is that it is straightforward to simulate empirical null distributions of complex statistics such as percentiles, odds ratios, or correlation coefficients. The exact distributions of OPLSR vectors are too complex to obtain, and so are even approximations to the distributions. Thus the permutation test is employed to test the significance of individual variables in the OPLSR vector.

To propose a permutation test procedure to test the significance of the $\hat{\beta}_{OPLS}$ coefficients, the basic procedures are presented as follows:

Step 1 Perform PLS with Y and X^{OSC}.
Step 2 Obtain the observed values of $\hat{\beta}_{OPLS,i}$ for all i.
Step 3 For each variable i, repeat the following a large number of times (e.g., 999 times):

Step 3.1 Randomize the values within the ith column of X^{OSC}.
Step 3.2 Perform PLS with Y and the permuted X^{OSC}.
Step 3.3 Obtain the realized value of $\hat{\beta}_{OPLS,i}$

Step 4 Estimate the p-value p_i for the ith variable (e.g., (the number of the realized equal to or larger than the observed in terms of absolute values+1)/1000).
Step 5 Choose a significance level α_{pm} so that variables in which $p_i < \alpha_{pm}$ are selected.

For the randomization procedure at Step 3.1, one can choose several columns of X randomly, i.e., a portion of the total number of variables in one iteration to speed up the procedure since the total number of variables is usually quite large. We note that the proportion of 0.3 worked well in practice, producing the empirical null distributions of realized $\hat{\beta}_{OPLS,i}$ that were highly close to normality. In this manner, the realized values of $\hat{\beta}_{OPLS,i}$ can be regarded as emerging from a collection of the null hypotheses that the variables effects are insignificant. The motivation is that practically we can assume a substantial number of variables are negligible among the large number of variables, which is quite common in spectral data. The coefficients $\hat{\beta}_{OPLS,i}$ for the variable i become zero in theory if the variable i does not contribute to the construction of response variables Y. Accordingly, the p-value is computed according to the extreme 'two-tailed' directions at Step 4. We also note that at Step 4, the observed value is included as one of the possible values of the randomization procedure. To control the increase of Type I error due to multiple testing, a correction can be made at Step 5. For instance, one can adjust the significance level α_{pm} to be α_{pm}/p using the Bonferroni procedure. In this study, we set α_{pm} to 0.05 due to its practicality. As a preprocessing step for the permutation test with significance level α_f, we filter out the variables of which the individual correlation measures with Y are weak. This step eliminates unnecessary noisy variables that can act as contributing variables from the whole procedure, so that only filtered-in variables are considered. The level of α_f is considered as a tolerance level for the OPLSR method and usually set to 0.05. By changing the level of α_f, we can control the number of variables included in the permutation test.

Simulation and comparisons

To test the OPLS regression-based feature selection method, we performed a simulation study. We first generated a data matrix $X(40 \times 1000)$ and a response matrix $Y(40 \times 1)$, comprising 40 samples and 1000 variables per sample. Each element of Y was a Bernoulli random variable with success probability 0.4, so 16 elements of Y were set to 1 on average while the remaining 24 elements were set to 0. To generate 1000 variables, we used a three-layer network structure as shown in Fig. 1a.

In the first layer, the first 30 variables were generated to have high separation in Y: for $p = 1, \ldots, 4, x_p \sim U(0,1) + 0.8 - 2Y$ and for $p = 5, \ldots, 9, x_p \sim U(0,1) - 1.2 - 2Y$, where $U(a,b)$ is a random variable from a uniform distribution of range a and b. This means the variables from x_1 to x_4 individually had a high positive correlation with Y (labeled 0), whereas x_5 to x_9 were highly correlated with Y (labeled 1). For instance, Fig. 1b shows the plots of realizations of x_1 and x_5 and the aforementioned pattern that individually separate Y. The first 30 variables in the first layer represent strong individual variables that can identify pathological conditions. For $p = 10, 13, \ldots, 28$,

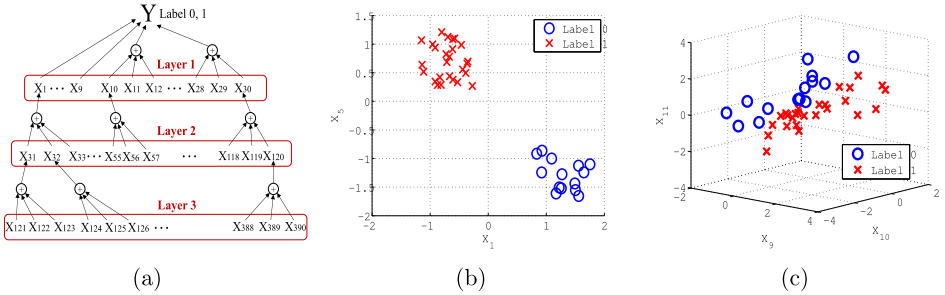

Fig. 1 **a** A data model to generate 1000 variables has three layers of variable interation that contribute to the response *Y*. Layer 1 contains 30 variables, including eight strong variables (x_1 to x_9) and the other group-wise variables (x_{10} to x_{30}). Each of the variables in Layer 1 is determined by three variables in Layer 2, and each variable in Layer 2 by three variables in Layer 3. The remaining 610 variables are randomly and independently assigned. **b** Two strong variables x_1 and x_5 separate the two labels of *Y*. **c** Three variables in Layer 1 jointly separate the two labels of response *Y*

$$\left[x_p\ x_{p+1}\ x_{p+2}\right]^T \sim N\left(0 + Y[1\ 2\ 3]^T, \begin{bmatrix} 12 & 10 & 8 \\ 10 & 12 & 10 \\ 8 & 10 & 12 \end{bmatrix}\right),$$

where $N(\mu, \Sigma)$ represents a multivariate normal distribution with mean μ and covariance Σ. Figure 1c shows plots of realizations of three variables in the layer, generated as above, jointly discriminate *Y*. The variables $x_i, i = 10, \ldots, 30$, in the first layer represent strong group-wise variables that strongly discriminate pathological conditions.

The next 90 variables from x_{31} to x_{120} in the second layer and the next 270 variables from x_{121} to x_{390} in the third layer were generated so that those variables contribute to the overall response *Y* in a composite and aggregate manner. For instance, x_1 in the first layer is clearly separable by a combination of x_{31}, x_{32}, and x_{33} in the second layer. Specifically, the generation of three variables is based on the value x_1 so that the sum of the three will be close to x_1 as follows: we set $x_{31} = \frac{u_1}{u_1+u_2+u_3}(x_1 + \epsilon), x_{32} = \frac{u_2}{u_1+u_2+u_3}(x_1 + \epsilon)$, and $x_{33} = \frac{u_3}{u_1+u_2+u_3}(x_1 + \epsilon)$ using independently generated u_1, u_2, and u_3 from $U(0, 1)$ and ϵ from $N(0, x_1/10)$. We notice that $x_{31} + x_{32} + x_{33} = x_1 + \epsilon$. The remaining 610 variables from x_{391} to x_{1000}, comprising a noise layer, were randomly and independently generated from $N(0, 1)$ so as to have a weak correlation with *Y*. They represent the presence of inherent noise. The adopted three-layers structure is a simulated example of multiple layers of spectral collinearity, interaction, and regulation in complex biological systems. For instance, a biological system for nutritional metabolomics reflects such a layered structure with linked transports [27].

Using the simulated data, we tested the performance of the proposed methods in comparison with the false discovery rate (FDR) method to detect the known variables in the three layers. Focusing on the effects of individual variables and controlling family-wise Type I error, FDR serves as a baseline method to compare the OPLSR method with. The two types of OPLS regression vector, $\hat{\beta}_{OPLSa}$ and $\hat{\beta}_{OPLSb}$, were considered. The number of predictive components for the OPLS regression model, *A* as in (6), varied from 1 to 3 so as to determine its effect upon the performance. The filtering level α_f for the OPLSR method and the *q*-value for FDR varied from 0.01 to 0.05 and 0.10. We repeated this test 1000 times, and in each repetition for each method, the number of variables that were

found within the three layers were counted. No variables among the random 610 variables were found for each of the two methods.

To illustrate the performance of the proposed method, Fig. 2 shows illustrations of the empirical null distribution of the realized $\hat{\beta}_{OPLSb,i}$ and regression vector $\hat{\beta}_{OPLSb}$. The empirical null distribution of $\hat{\beta}_{OPLSb,i}$ in Fig. 2a, obtained by the permutation procedure, is closely normally distributed and provides a basis for testing the observed $\hat{\beta}_{OPLSb,i}$. Figure 2b also shows the empirical null distribution of the realized $\hat{\beta}_{OPLSb}$ for the variable with the largest confidence interval. This illustration, being the worst case, demonstrates that the distribution of the realized regression coefficients for individual variables can be sufficiently approximated to a normal distribution by the setting. The regression vector $\hat{\beta}_{OPLSb}$ in Fig. 2c highlights the filtered-in variables according to the α_f-level filtering

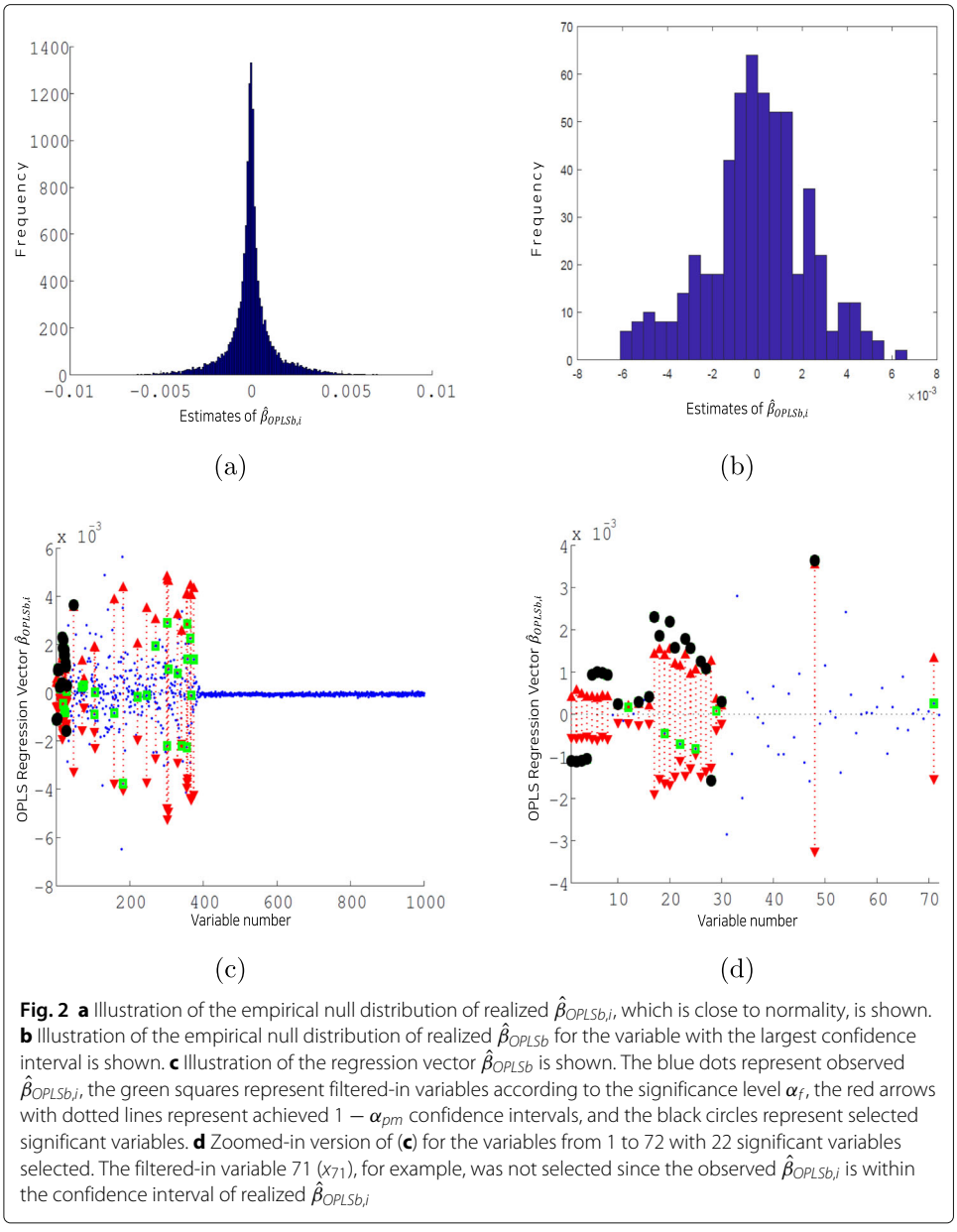

Fig. 2 **a** Illustration of the empirical null distribution of realized $\hat{\beta}_{OPLSb,i}$, which is close to normality, is shown. **b** Illustration of the empirical null distribution of realized $\hat{\beta}_{OPLSb}$ for the variable with the largest confidence interval is shown. **c** Illustration of the regression vector $\hat{\beta}_{OPLSb}$ is shown. The blue dots represent observed $\hat{\beta}_{OPLSb,i}$, the green squares represent filtered-in variables according to the significance level α_f, the red arrows with dotted lines represent achieved $1 - \alpha_{pm}$ confidence intervals, and the black circles represent selected significant variables. **d** Zoomed-in version of (**c**) for the variables from 1 to 72 with 22 significant variables selected. The filtered-in variable 71 (x_{71}), for example, was not selected since the observed $\hat{\beta}_{OPLSb,i}$ is within the confidence interval of realized $\hat{\beta}_{OPLSb,i}$

(green squares), confidence intervals of the permutation procedure (red arrows with dotted lines), and selected variables (black circles) along with the observed regression vector $\hat{\beta}_{OPLSb,i}$ (blue dots). We notice that observed $\hat{\beta}_{OPLSb,i}$ for $i = 1, \cdots, 390$ are substantially larger than those for the rest, which indicates the regression vector reflects the network structure for the data set. The selected variables gather in the first layer as shown in Fig. 2d, which implies the testing procedure of the method suitably separate known important variables. Particularly, we notice that the variables (from x_1 to x_8) are positioned accordingly with their contributions to the labels. For example, the realized value $\hat{\beta}_{OPLSb,i}$ for x_1 is negative, implying that the increase of x_1 results in the increase of label 0 instead of label 1. This is in accordance with the behavior of x_1 in Fig. 1b. In Fig. 2d we also observe that the filtered-in variable 71 (x_{71}), for example, was not selected since the observed $\hat{\beta}_{OPLSb,i}$ is within the confidence interval of the realized $\hat{\beta}_{OPLSb,i}$.

Additionally, Table 1 shows the amounts of variation of X^{ORTH}, X^{OSC}, and Y in the OPLSR method. The first OSC component and the first PLS component accounted for more than 96% of the total X variation and 99% of the total X^{OSC}, respectively. It is an empirical evidence that using the first OSC component and the first one or two PLS components is sufficient. The first PLS component accounted for more than 99% of the Y variation in the simulation study. In fact, we used the first OSC component when correcting X in (5). Table 2 shows the average numbers of selected variables for OPLSR and FDR according to significance level α. Since the variables in layer 1 are significant, the selected variables in layer 1 mean the recall rate. The OPLSR method is divided into $\hat{\beta}_{OPLSa}$ (denoted by OPLSR$_a$) and $\hat{\beta}_{OPLSb}$ (denoted by OPLSR$_b$). We chose significance level α as q-value for FDR and α_f for OPLSR. We note that the levels for the methods are not strictly equivalent by themselves, yet we compare them in that they practically adjust the number of selected variables.

The results for this simulation study show that OPLSR$_a$ consistently found more variables than FDR. The performance of OPLSR$_b$, working quite well, was comparable with that of FDR while surpassing FDR for big α.

Experiments

We applied the proposed method in the examination of a high-resolution metabolomics data set. We show the use of the proposed method, and furthermore analysis will follow with the next dataset. The code and data are uploaded to the following GitHub url: https://github.com/leegs52/OPLSR. The metabolomics data set consists of 127 samples screened for bile acids by a BioQuant colormetric kit [28]: 64 samples have bile acids present and

Table 1 Amounts of variation of X^{ORTH}, X^{OSC}, and Y in the OPLSR method for the simulation study are shown. The first OSC component and the first PLS component accounted for more than 96% of the total X variation and 99% of the total X^{OSC}, respectively. The first PLS component accounted for more than 99% of the Y variation in the simulation study

			PLS					PLS		
	OSC		1	2	3	OSC		1	2	3
X^{ORTH}		0.967					0.967			
X^{OSC}	1	0.033	0.991	0.002	0.002	2	0.033	0.994	0.002	0.001
Y		1.00	1.00	0.000	0.000		1.00	1.00	0.000	0.000

Table 2 The average numbers of selected variables in the layers for each method of OPLSR, according to $\hat{\beta}_{OPLSa}$ (denoted by OPLSR$_a$) and $\hat{\beta}_{OPLSb}$ (denoted by OPLSR$_b$), and FDR; and significance level α, meaning α_f for OPLSR$_b$ and q-value for FDR, are shown. This shows that OPLSR$_a$ found more variables consistently while OPLSR$_b$ worked comparably to FDR

α	0.01			0.05			0.10		
	Layer			Layer			Layer		
	1	2	3	1	2	3	1	2	3
OPLSR$_a$	24.4	0.837	0.580	25.0	3.79	5.61	25.1	5.32	10.0
OPLSR$_b$	18.1	0.190	0.137	20.8	0.987	1.24	22.9	2.85	3.68
FDR	21.3	0.020	0.000	22.1	0.100	0.027	22.8	0.103	0.010

the others have bile acids absent. The metabolomic profiling was performed using liquid chromatography (LC) coupled to high-resolution mass spectrometry by an Orbitrap FTQ-Velos mass spectrometer. Peak extraction and quantification of ion intensities were performed by an adaptive processing of liquid chromatography mass spectroscopy software package that produced 7068 m/z (mass divided by charge number of ions) values. We aimed to extract significant metabolites, for example the top 1%, that separate bile acid present and bile acid absent. Application of the OPLSR method to the data matrix X of size 127×7068 and response Y of size 127×1 yielded the empirical null distribution of OPLS regression vector $\hat{\beta}_{OPLSb}$. Figure 3 shows two-sided 95% confidence intervals (red rectangles) and significant m/z variables (black circles) using the permutation testing of OPLS regression. The small blue dots represent the OPLS regression coefficients. The proposed method was able to filter out significant variables in that the number of found significant variables (black circles) by the OPLS is 11 while the number of variables

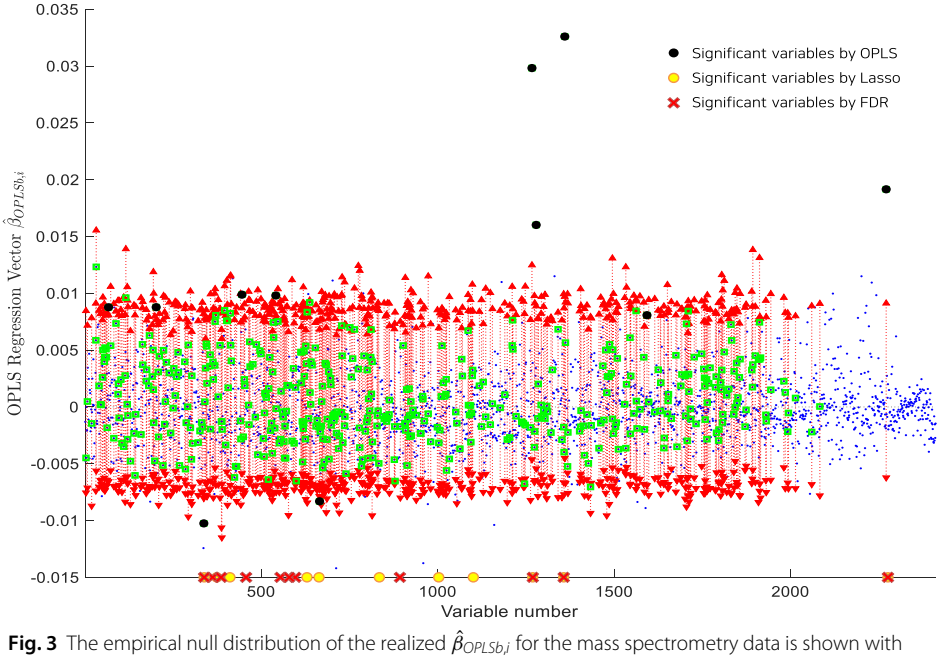

Fig. 3 The empirical null distribution of the realized $\hat{\beta}_{OPLSb,i}$ for the mass spectrometry data is shown with two-sided 95% confidence intervals by red rectangles. The black dots mean significant m/z variables by the permutation testing of OPLS regression. The yellow dots and the red crosses illustrate siginificant variables by the Lasso and the FDR, respectively

(green circles) by individual coefficient testing of regression analysis is 461. The number of significant variables by either Lasso (marked by yellow circles) or FDR (marked by red crosses) are 11, like the OPLS. For information, all three of them found x_{1267}, x_{1360}, and x_{2271} the significant variables.

We also applied the proposed method to a near-infrared (NIR) spectroscopic technique, which is a useful tool for chemical process analysis in research such as pharmaceutical, medical diagnostics, and agrochemical quality. Measured NIR spectroscopic spectra are influenced by external process variables such as temperature, pressure, and viscosity. The difficulty in keeping these variables unchanging and the necessity to change their value during the process (e.g., setting temperature and pressure in batch processes) make it necessary to assess the influence on the NIR spectra. Wülfert et al. took short-wave NIR spectra of ethanol, water, and 2-propanol mixtures at different temperatures to assess the influence of those temperature-induced spectra variations [29]. The proposed method was applied to the 22 spectra of the mixtures at each temperature of 30, 40, 50, 60, and 70 °C ($n = 22 \times 5 = 110$). The used wavelengths are from 580 to 1091 nm, resulting 512 variables ($p = 512$). The overall 110 spectra are shown in Fig. 4a.

The data set consists of the 110 spectra as X (110×512) and temperatures as Y (110×1). For the purpose of testing predictive power, the data set was randomly split into a training set with 70% of the whole for building a prediction model and a test set with the remaining 30% for estimating the predictive quality of that model. Using the selected variables for each method of OPLSR$_a$, OPLSR$_b$, FDR, and Lasso, we carried out regression analysis to predict Y of the test set and calculated the mean-squared error (MSE) of prediction and the respective amount of Y variance being described by the model, Q^2, as follows:

$$Q^2 = 1 - \frac{\sum_i (Y_{i,test} - \hat{Y}_{i,test})^2}{\sum_i Y_{i,test}^2}.$$

This was repeated 3,000 times. We also measured the precision, denoted by `P`, and the number of selected variables, denoted by `N`, as additional information on the performance of each method. The precision `P` is the fraction of correctly selected variables compared to all selected variables. For the NIR spectra of temperatures, Fig. 4a shows that the variation between them is highly significant around 980 nm with a peak maximum [30]. By

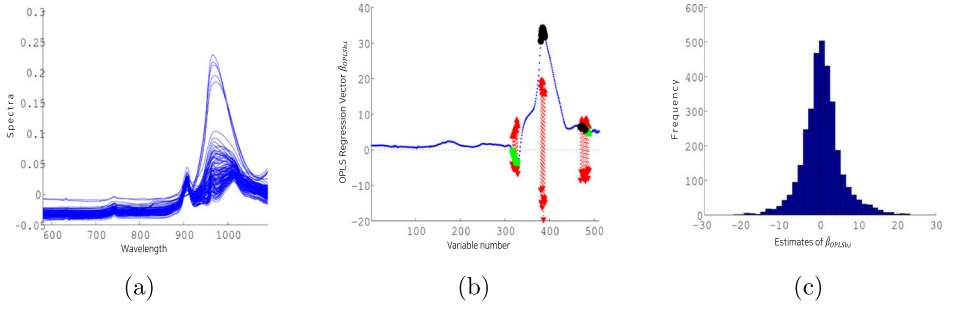

Fig. 4 (**a**) Overall 110 NIR spectra for temperatures of 30, 40, 50, 60, and 70 °C are shown. Illustration of (**b**) the regression vector $\hat{\beta}_{OPLSb}$ and (**c**) the empirical null distribution of the realized $\hat{\beta}_{OPLSb,i}$ are shown for the NIR spectra. The wavelength ranges of the selected variables were from 960 to 970 nm (x_{381} to x_{391}) and from 1049 to 1060 nm (x_{470} to x_{481})

considering each wavelength within the 930-1060 nm as the ground-truth features, we verified the selected variables.

The illustrations of the proposed method for the NIR spectra are shown in Fig. 4b and c, which depict the regression vector $\hat{\beta}_{OPLSb}$ and the empirical null distribution of realized $\hat{\beta}_{OPLSb,i}$, respectively. The wavelength ranges of the selected variables were from 960 to 970 nm (x_{381} to x_{391}), which corresponds to the free OH 2nd overtone, and from 1049 to 1060 nm (x_{470} to x_{481}), which belongs to the hydrogen-bonded OH groups [31]. This finding is consistent with the chemical observation that temperature effects are related to the absorbance of molecule overtones and the cluster size of hydrogen-bonded molecules [29]. The empirical null distribution in Fig. 4c, following closely normality, supply the proposed method with the basis of the employed permutation test.

Table 3 shows the comparison results in terms of evaluation metrics and the number of selected variables. We observed that the OPLSR method reliably finds significant variables among the spectra most of the iterations since its precision outperformed FDR and Lasso. It is not surprising because the proposed method examines the effects of variables in a collective manner by considering the covariance structure and the effects of both orthogonal and predictive components at the same time. The regression performance for the variables selected by Lasso achieved relatively high predictive performance. Though Lasso was unsatisfactory in the precision. For all criteria of the precision, MSE, and Q^2, OPLSR$_b$ outperformed FDR consistently. The approach OPLSR$_a$ also outperformed FDR consistently. The method OPLSR$_b$ outperformed OPLSR$_a$ for $\alpha_f = 0.01, 0.05$ in terms of the precision, MSE, and Q^2, while there was little difference between the two for $\alpha_f = 0.10$.

Additionally, in order to demonstrate the robustness of the proposed method, we conducted down-sampling and then performed another experiment under the same conditions of Table 3. The down-sampled data set was randomly drawn from NIR spectra data ($n = 110$), and it contains 80 samples, i.e., about 70 percent of total. For the down-sampled data, Table 4 presents the performance comparison for the four methods, demonstrating the degree of robustness.

Table 3 Performance of each method of OPLSR$_a$, OPLSR$_b$, FDR, and Lasso is shown with respect to each parameter (q-value for FDR, λ for Lasso, and α_f for the other two). The precision and the number of selected variables were denoted by P and N, respectively. The boldfaced numbers indicate the ones which outperformed the others

Method		MSE	Q^2	P	N
OPLSR$_a$	α_f=0.01	12,332	0.86	0.871	11.0
	α_f=0.05	8,989	0.89	0.701	16.2
	α_f=0.10	**6,829**	0.91	0.640	22.7
OPLSR$_b$	α_f=0.01	10,896	0.88	**0.915**	11.7
	α_f=0.05	8,704	0.91	0.860	28.3
	α_f=0.10	7,081	**0.92**	0.818	41.2
FDR	q=0.01	33,038	0.64	0.750	9.0
	q=0.05	12,123	0.86	0.714	36.6
	q=0.10	11,194	0.85	0.536	40.1
Lasso	λ=0.01	7,120	0.91	0.545	6.6
	λ=0.10	10,796	0.89	0.477	4.0
	λ=0.50	14,028	0.81	0.496	2.2

Table 4 Performance of each method on downsampled data. The boldfaced numbers indicate the ones which outperformed the others

Method		MSE	Q^2	P	N
$OPLSR_a$	α_f=0.01	10,972	0.83	0.883	8.90
	α_f=0.05	9,105	0.87	0.777	24.4
	α_f=0.10	**5,709**	0.91	0.658	22.8
$OPLSR_b$	α_f=0.01	10,133	0.85	**0.884**	10.4
	α_f=0.05	8,811	0.89	0.881	31.0
	α_f=0.10	7,782	**0.93**	0.848	48.2
FDR	q=0.01	23,116	0.65	0.701	6.0
	q=0.05	10,330	0.85	0.685	39.5
	q=0.10	13,194	0.85	0.554	40.0
Lasso	λ=0.01	6,235	0.90	0.502	6.3
	λ=0.10	9,886	0.88	0.494	3.2
	λ=0.50	13,124	0.80	0.498	2.1

Conclusions

We presented a feature selection method based on orthogonal-signal corrected PLS regression vectors to identify significant variables associated with the response characteristics. The proposed OPLSR method integrates PLS with orthogonal signal correction and permutation tests. To remove unnecessary variation in the input variables and improve interpretability of PLS regression vectors, orthogonal signal correction was applied first. The orthogonal-signal corrected PLS procedure reflects the variable interrelationships under complex systems, which are easily represented in a network structure. The two types of regression vectors from the model, carrying information on variables contribution to response characteristics, were derived and investigated in both a simulation study and a real-life spectra study in contrast to FDR and Lasso. To select the significant variables from the regression vectors, we applied a permutation test that generates empirical null distributions of variable effects on the response characteristics. The adopted permutation test was provided with the filtering rate, a pre-defined tolerance level for the whole selection procedure for eliminating unnecessary noisy variables, and was implemented efficiently by taking advantage of a collection of insignificant variables. Through simulations, we observed that the proposed method well captured the predefined network structures and successfully found the known variables. We demonstrated this method with real-life metabolomics and NIR spectra data, the finding variables that achieve a good level of predictive power and accurately relate to the chemical observations. For future research, we hope to investigate the effect of imbalance classes in the feature selection.

Acknowledgments
Not applicable.

Conflict of interest
None to declare.

Authors' contributions
GL performed analyses on the simulated data and participated in the writing of the manuscript. KL implemented the algorithms, motivated the research problem, and designed the study. All authors read and approved the final manuscript.

Funding
None to declare

Availability of data and materials
The implemented MATLAB package is available from the following url address: https://github.com/leegs52/OPLSR

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 June 2020 Accepted: 10 January 2021
Published online: 22 January 2021

References

1. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70–9.
2. Zhu Z, Ong Y-S. Memetic algorithms for feature selection on microarray data. In: International Symposium on Neural Networks. Berlin Heidelberg: Springer-Verlag; 2007. p. 1327–35.
3. Alsberg BK, Woodward AM, Winson MK, Rowland JJ, Kell DB. Variable selection in wavelet regression models. Anal Chim Acta. 1998;368(1-2):29–44.
4. Joe Qin S. Statistical process monitoring: basics and beyond. J Chemometr. 2003;17(8-9):480–502.
5. Holmes E, Antti H. Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst. 2002;127(12):1549–57.
6. Lindon JC, Holmes E, Nicholson JK. Pattern recognition methods and applications in biomedical magnetic resonance. Prog Nucl Magn Reson Spectrosc. 2001;39(1):1–40.
7. Park YH, Kong T, Roede JR, Jones DP, Lee K. A biplot correlation range for group-wise metabolite selection in mass spectrometry. BioData Min. 2019;12(1):4.
8. Gerlach RW, Kowalski BR, Wold HOA. Partial least-squares path modelling with latent variables. Anal Chim Acta. 1979;112(4):417–21.
9. Brás LP, Lopes M, Ferreira AP, Menezes JC. A bootstrap-based strategy for spectral interval selection in PLS regression. J Chemometr. 2008;22(11-12):695–700.
10. Jiang J-H, Berry RJ, Siesler HW, Ozaki Y. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. Anal Chem. 2002;74(14):3555–65.
11. Heise HM, Bittner A. Rapid and reliable spectral variable selection for statistical calibrations based on PLS-regression vector choices. Fresenius J Anal Chem. 1997;359(1):93–9.
12. Høskuldsson A. Variable and subset selection in PLS regression. Chemometr Intell Lab Syst. 2001;55(1-2):23–38.
13. Faber NKM. Uncertainty estimation for multivariate regression coefficients. Chemometr Intell Lab Syst. 2002;64(2):169–79.
14. Wehrens R, Van der Linden WE. Bootstrapping principal component regression models. J Chemometr Soc. 1997;11(2):157–71.
15. Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. Chemometr Intell Lab Syst. 1998;44(1-2):175–85.
16. Fearn T. On orthogonal signal correction. Chemometr Intell Lab Syst. 2000;50(1):47–52.
17. Svensson O, Kourti T, MacGregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. J Chemometr. 2002;16(4):176–88.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
19. Kim SB, Chen VCP, Park Y, Ziegler TR, Jones DP. Controlling the false discovery rate for feature selection in high-resolution NMR spectra. Stat Anal Data Min. 2008;1(2):57–66.
20. Vinzi VE, Chin WW, Henseler J, Wang H, et al. Handbook of partial least squares, vol 201. Berlin Heidelberg: Springer-Verlag; 2010.
21. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemometr Soc. 2002;16(3):119–28.
22. Westerhuis JA, de Jong S, Smilde AK. Direct orthogonal signal correction. Chemometr Intell Lab Syst. 2001;56(1):13–25.
23. de Jong S. SIMPLS: an alternative approach to partial least squares regression. Chemometr Intell Lab Syst. 1993;18(3):251–63.
24. Efron B, Tibshirani R. An introduction to the bootstrap, vol 57. Boca Raton: CHAPMAN & HALL/CRC CRC Press LLC; 1993.
25. Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat Med. 2019;38(4):558–82.
26. Tsuzuki S, Fujitsuka N, Horiuchi K, Ijichi S, Gu Y, Fujitomo Y, Takahashi R, Ohmagari N. Factors associated with sufficient knowledge of antibiotics and antimicrobial resistance in the Japanese general population. Sci Rep. 2020;10(1):1–9.
27. Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: progress in addressing complexity in diet and health. Annu Rev Nutr. 2012;32:183–202.

28. Neujahr DC, Uppal K, Force SD, Fernandez F, Lawrence C, Pickens A, Bag R, Lockard C, Kirk AD, Tran V, et al. Bile acid aspiration associated with lung chemical profile linked to other biomarkers of injury after lung transplantation. Am J Transplant. 2014;14(4):841–8.
29. Wülfert F, Kok WT, Smilde AK. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. Anal Chem. 1998;70(9):1761–7.
30. Zhang G, Cui Q, Liu G. Efficient near-infrared quantum cutting and downshift in Ce3+-Pr3+ codoped SrLaGa3S6O suitable for solar spectral converter. Opt Mater. 2016;53:214–7.
31. Bonanno AS, Olinger JM, Griffiths PR. The origin of band positions and widths in near infrared spectra. Near Infrared Spectroscopy: bridging the gap between data analysis and NIR applications. London: Ellis Horwood. 1992;19–28.

Publisher's Note