

Integration and variable selection of ‘omics’ data sets with PLS: a survey

Titre: Une revue sur l’intégration et la sélection de variables ‘omiques’ avec la PLS

Kim-Anh Lê Cao¹ and Caroline Le Gall²

Abstract: ‘Omics’ data now form a core part of systems biology by enabling researchers to understand the integrated functions of a living organism. The integrative analysis of these transcriptomics, proteomics, metabolomics data that are co jointly measured on the same samples represent analytical challenges for the statistician to extract meaningful information and to circumvent the high dimension, the noisiness and the multicollinearity characteristics of these multiple data sets. In order to correctly answer the biological questions, appropriate statistical methodologies have to be used to take into account the relationships between the different functional levels. The now well known multivariate projections approaches greatly facilitate the understanding of complex data structures. In particular, PLS-based methods can address a variety of problems and provide valuable graphical outputs. These approaches are therefore an indispensable and versatile tool in the statistician’s repertoire.

Variable selection on high throughput biological data becomes inevitable to select relevant information and to propose a parsimonious model. In this article, we give a general survey on PLS before focusing on the latest developments of PLS for variable selection to deal with large omics data sets. In a specific discriminant analysis framework, we compare two variants of PLS for variable selection on a biological data set: a backward PLS based on Variable Importance in Projection (VIP) which good performances have already been demonstrated, and a recently developed sparse PLS (sPLS) based on Lasso penalization of the loading vectors.

We demonstrate the good generalization performance of sPLS, its superiority in terms of computational efficiency and underline the importance of the graphical outputs resulting from sPLS to facilitate the biological interpretation of the results.

Résumé : Les données ‘Omiques’ sont largement utilisées en biologie des systèmes pour comprendre les mécanismes biologiques impliqués dans le fonctionnement des organismes vivants. L’intégration de ces données transcriptomiques, protéomiques ou métabolomiques parfois mesurées sur les mêmes échantillons représente un challenge pour le statisticien. Il doit être capable d’extraire de ces données les informations pertinentes qu’elles contiennent, tout en devant composer avec des données à grandes dimensions et souffrant fréquemment de multicollinéarité. Dans ce contexte, il est primordial d’identifier les méthodes statistiques capables de répondre correctement aux questions biologiques, mêlant parfois des relations entre différents niveaux de fonctionnalité. Les techniques statistiques multivariées de projections dans des espaces réduits facilitent grandement la compréhension des structures complexes des données omiques. En particulier, les approches basées sur la méthode PLS constituent un outil indispensable à la panoplie du statisticien. Leur grande polyvalence permet d’adresser une large variété de problèmes biologiques tout en fournissant des résultats graphiques pertinents pour l’interprétation biologique.

Etant donné le grand nombre de variables considérées (gènes, protéines ...), la sélection de variables est devenue une étape inévitable. L’objectif est de sélectionner uniquement l’information pertinente afin de construire le modèle le plus parcimonieux possible. Dans cet article, nous présentons la méthode PLS puis nous mettons l’accent sur les derniers développements en matière de sélection de variables pour la PLS dans le cadre de données omiques abondantes. Deux approches de sélection de variables avec PLS sont comparées dans le cas d’une analyse discriminante appliquée à un jeu de données biologiques : une approche descendante (‘backward’) basée sur le critère du VIP (‘Variable Importance

¹ Queensland Facility for Advanced Bioinformatics, University of Queensland, 4072 St Lucia, QLD, Australia.
E-mail: k.lecao@uq.edu.au

² Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France.
E-mail: Caroline.Le-Gall@insa-toulouse.fr

in Projection') pour laquelle de bonnes performances ont déjà été démontrées dans la littérature et la sparse PLS (sPLS), une approche récente basée sur une pénalisation Lasso des vecteurs 'loadings'.

La sparse PLS montre de très bonnes performances globales ainsi qu'une très nette supériorité en temps de calcul. Elle permet aussi de démontrer l'efficacité des représentations graphiques issues de la PLS dans l'interprétation biologique des résultats.

Keywords: Partial Least Squares regression, variable selection

Mots-clés : régression Partial Least Squares, sélection de variables

AMS 2000 subject classifications: 6207, 62H99, 62P10, 62H30

Introduction

Challenges when $n \ll p$ and variable selection. Each omics platform is now able to generate a large amount of data. Genomics, proteomics, metabonomics/metabolomics, interactomics are compiled at an ever increasing pace and now form a core part of the fundamental systems biology framework. These data are required to understand the integrated functions of the living organism. However, the abundance of data is not a guarantee of obtaining useful information in the investigated system if the data are not properly processed and analyzed to highlight this useful information.

From a statistical point of view, the goodness of a model is often defined in terms of prediction accuracy - for example in a regression framework. However, parsimony is crucial when the number of predictors is large, as most statistical approaches predict poorly because of the noisiness and the multicollinearity characteristics of the data. Simpler and sparse models with few covariates are preferred for a better interpretation of the model, a better prediction of the response variable, as well as a better understanding of the relationship between the response and the covariates.

A variety of biological questions. A major challenge with the integration of omics data is the extraction of discernible biological meaning from multiple omics data. It involves the identification of patterns in the observed quantities of the dynamic intercellular molecules (mRNAs, proteins, metabolites) in order to characterize all the elements that are at work during specific biological processes. Studying biology at the system level enables (a) to identify potential functional annotation, for instance, to assign specific enzymes to previously uncharacterized metabolic reactions when integrating genomics and metabolomics data, (b) to identify biomarkers associated with disease states and elucidate signalling pathway components more fully, for instance to define prognosis characteristics in human cancers by using transcriptomics signatures to identify activated portions of the metabolic network (c) to address fundamental evolutionary questions, such as identifying cellular factors that distinguish species; these factors likely have had roles in speciation events (d) to interpret toxicological studies (toxicogenomics) or (e) to study the complex reactions between the human body, nutritional intake and the environment (nutrigenomics). Many other central questions can be addressed with omics data integration. These data may not be sufficient to understand all the underlying principles that govern the functions of biological systems but they will nonetheless allow investigators to tackle difficult problems on previously unprecedented scales.

A variety of statistical approaches. Many statistical approaches can be used to analyse omics data. We list some of them and discuss why multivariate projection methods might be adequate to

give more insight into omics data sets and, ultimately, to enable a more fundamental understanding of biology.

Univariate analysis. Univariate analysis has been extensively used in microarray analysis to look for the expression of differentially expressed genes. However, it does not take into account the relations between the variables. Multi-variable patterns can be significant, even if the individual variables are not. Most importantly, in the case of data integration in systems biology, it is absolutely crucial to take into account the relationships between the different omics data.

Machine learning approaches. Machine learning approaches take into account the correlation between the variables but are often considered as black boxes. They also often require dimensionality reduction and are extremely computationally demanding.

Network analyses. The inference of networks is of biological importance and is intrinsically linked to data integration. It provides useful outputs to visualize the correlation structure between the omics data sets and allows to check/propose new hypotheses on biological pathways.

Multivariate projection methods. In the omics era, data-driven analysis by means of multivariate projection methods greatly facilitates the understanding of complex data structures. The advantages of multivariate projection methods is their application to almost any type of data matrix, e.g. matrices with many variables, many observations, or both. Their flexibility, versatility and scalability make latent variable projection methods particularly apt at handling the data-analytical challenges arising from omics data, and they can effectively handle the hugely multivariate nature of such data. They produce score vectors and weighted combinations of the original variables that enable a better insight into the biological system under study.

Multivariate projection methods, such as PLS-based methods are seen as remarkably simple approaches, and they often have been overlooked by statisticians as it has been considered as an algorithm rather than a rigorous statistical model. Yet within the last years, interest in the statistical properties of PLS has risen. PLS has been theoretically studied in terms of its variance and shrinkage properties [37, 22, 8, 24]. The literature regarding PLS methods is very extensive. The reader can refer to the reviews of [49, 52, 38, 7]. PLS is now seen as having a great potential to analyse omics data, not only because of its flexibility and the variety of biological questions it can address, but also because its subsequent graphical outputs allow to interpret the results [16, 30, 17]. In particular, a variant called O2-PLS has been extensively used in metabonomics data [9] but will not be presented in this review.

In this review. In Section 1, we first survey variants of PLS for the integration of two omics data sets and present different analysis frameworks. In Section 2, we then particularly emphasize on variable selection with PLS for large biological data sets in order to select the relevant information and remove noisy variables. Extremely valuable by-products of PLS-based methods are the graphical outputs which facilitate the interpretation of the results and give a better understanding of the data. In Subsection 3.1, on a biological data that include transcripts and clinical variables, we illustrate how these graphical outputs can help give more insight into the biological study. In Subsection 3.2 and in a classification framework, we numerically compare two PLS variants for variable selection on the transcriptomics data: the first variant is a backward approach based on VIP, which good performances have been demonstrated by [11]; the second variant, sparse PLS-Discriminant Analysis (sPLS-DA) was recently developed by [31, 28] and includes Lasso penalization [44] on the loading vectors to perform variable selection. We show

the good generalization performance of sPLS-DA and its superiority in terms of computational efficiency. In Section 4, we finally discuss the validation of PLS results for the integration of complex biological data sets characterized by a very small number of samples.

1. Integrating two data sets with PLS

PLS is a multivariate method that enables the integration of two large data sets. In this section we present the Partial Least Square regression (PLS) algorithm and explain why PLS is efficient in a highly dimensional setting.

1.1. Notations

Throughout the article, we will use the following notations: the two matrices X and Y are of size $n \times p$ and $n \times q$ and form the training data, where n is the number of samples, or cases, and p and q are respectively the number of variables in X and Y . We will first present the general regression framework case of PLS2 where the response Y is a matrix - that also includes the regression (PLS1) for which $q = 1$. We will then focus on two other framework analyses that are special cases of PLS2: PLS-canonical mode uses a different deflation of the matrices to model a symmetric or bidirectional relationship between the two data sets and PLS-Discriminant Analysis deals with classification problems by coding Y as a dummy matrix. In this Section, we will primarily focus on a general PLS algorithm where $q \geq 1$.

Note that X and Y should be matched data sets, i.e. the two types of variables are measured on the same samples.

1.2. PLS regression

Introduction on PLS. Partial Least Squares regression (PLS) can be considered as a generalization of multiple linear regression (MLR). It relates two data matrices X and Y by a multivariate model, but it goes beyond traditional multiple regression in that it also models the structure of X and Y . Unlike MLR, PLS has the valuable ability to analyze many, noisy, collinear and even incomplete variables in both X and Y , and simultaneously models several response variables Y . Our regression problem is to model one of several dependent variables - or responses, Y by means of a set of predictor variables X . Example in genomics, if we consider the biology dogma, includes relating $Y =$ expression in metabolites to $X =$ expression of transcripts. The modelling of Y by means of X is traditionally performed using MLR, which works well as long as the X - variables are in small number and fairly uncorrelated, i.e. X is of full rank.

PLS therefore allow us to consider more complex problems. Given the deluge of data we are facing in genomics, it allows us to analyze available data in a more realistic way. However, the reader should keep in mind that we are far from a good understanding of the complications of biological systems and the quantitative multivariate analysis is still in its infancy, in particular with many variables and few samples.

The PLS algorithm. In PLS, the components called *latent variables* are linear combinations of the initial variables. However, the coefficients that define these components are not linear,

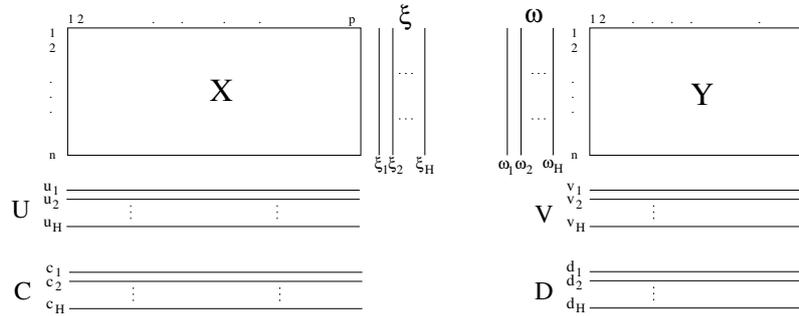


FIGURE 1. PLS scheme: the data sets X and Y are successively decomposed into sets of latent variables (ξ_1, \dots, ξ_H) , $(\omega_1, \dots, \omega_H)$ and loading vectors (u_1, \dots, u_H) , (v_1, \dots, v_H) . The vectors (c_1, \dots, c_H) and (d_1, \dots, d_H) are the partial regression coefficients and H is the number of deflations or dimensions in the PLS algorithm.

as they are solved via successive local regressions on the latent variables. The data sets X and Y are simultaneously modelled by successive decompositions. The objective function involves maximizing the covariance between each linear combination of the variables from both data sets:

$$\arg \max_{\|u_h\|=1, \|v_h\|=1} \text{cov}(Xu_h, Yv_h) \quad h = 1 \dots H. \quad (1)$$

The *loading vectors* are the vectors u_h and v_h for each PLS dimension h (H is the number of deflations), and the associated latent variables are denoted $\xi_h = Xu_h$ and $\omega_h = Yv_h$. The loading vectors u_h and v_h are directly interpretable, as they indicate how the variables from both data sets can explain the relationships between X and Y . The latent variables ξ_h and ω_h contain the information regarding the similarities or dissimilarities between individuals or samples.

In the following we present the PLS algorithm for the first deflation $h = 1$ (see also Fig. 1). Start: set ω to the first column of Y

1. $u = X^T \omega / \omega^T \omega$, scale u to be of length one. u is the *loading vector* associated to X
2. $\xi = Xu$ is the *latent variable* associated to X
3. $v = Y^T \xi / (\xi^T \xi)$, scale v to be of length one. v is the *loading vector* associated to Y
4. $\omega = Y^T v / (v^T v)$ is the latent variable associated to Y
5. If convergence then 6 else 1
6. $c = X^T \xi / \xi^T \xi$, $d = Y^T \omega / \omega^T \omega$ are the partial regression coefficients from the regression of X (Y) onto ξ (ω)
7. Compute the residual matrices $X \rightarrow X - \xi c^T$ and $Y \rightarrow Y - \omega d^T$

Step 6 performs local regressions of X and Y onto ξ and ω . By using successive local regressions on the latent vectors, the PLS algorithm therefore avoids the computation of the inverse of covariance or correlation matrix that might be singular.

The next set of iterations starts with the new X and Y residual matrices from previous iteration 7 (*deflation step*). The iterations can continue until a stopping criterion is used, such as the number of *dimensions* - chosen by the user, or if X becomes the zero matrix.

Note that the PLS algorithm actually performs in a similar way to the power iteration method of determining the largest eigenvalue for a matrix and will converge rapidly in almost all practical cases (less than 10 iterations).

The underlying model of PLS. We can write the multiple regression model of PLS as [14]:

$$X = \Xi C' + \varepsilon_1 \quad Y = \Xi D' + \varepsilon_2 = X\beta + \varepsilon_2,$$

where Ξ is the $(n \times H)$ column matrix of the latent variables ξ_h , and β ($p \times H$) is the coefficient regression matrix. The column matrices C and D are defined such that $c_h = X'_{h-1} \xi_h / (\xi'_h \xi_h)$ and $d_h = Y'_{h-1} \xi_h / (\xi'_h \xi_h)$, and ε_1 ($n \times p$) and ε_2 ($n \times q$) are the residual matrices, $h = 1 \dots H$. An insightful explanation on the geometric interpretation of PLS can be found in the review of [52].

Data scaling. The results of PLS or any projection method depend on the scaling of the data. In the absence of knowledge about the relative importance of the variables, the standard approach is to center each variable and scale them to unit variables. This corresponds to giving each variable (column) the same weight in the analysis.

1.3. PLS to highlight correlations

While PLS2 models an *asymmetric* or uni-directional relationship between the two data matrices, PLS-canonical mode can model a *symmetric* or bi-directional relationship. It can be used to predict Y from X and X from Y . For example, [30] applied this variant in a case where the same samples were measured using two different types of transcriptomics platforms to highlight correlated transcripts across the two platforms. This variant is particularly useful as an alternative to Canonical Correlation Analysis (CCA), which is limited by the number of variables leading to singular correlation matrices and ill-posed matrix problems.

In step 7 of the PLS algorithm, the data matrices can be symmetrically deflated with respect to each latent variable (see [49] for a detailed review). This deflation mode has been called *canonical* ([43], also called "PLS-mode A"), where the two matrices are deflated as follows:

$$\begin{aligned} c &= X^T \xi / \xi^T \xi & e &= Y^T \omega / \omega^T \omega \\ X &\rightarrow X - \xi c^T & Y &\rightarrow Y - \omega e^T \end{aligned}$$

When analyzing standardized data sets, [43] showed that PLS-canonical mode and CCA gave different, although similar results when $n < p + q$.

In following Section 2.2.3, we present several sparse variants of PLS-canonical mode that have been recently proposed in the literature.

1.4. PLS-Discriminant Analysis

Although PLS is principally designed for regression problems, it performs well for classification and discrimination problems, and has often been used for that purpose [4, 35, 42]. PLS-Discriminant Analysis (PLS-DA) is a special case of PLS2 and can be seen as an alternative to Linear Discriminant Analysis (LDA). LDA has often been shown to produce the best classification

results, but faces numerical limitations when dealing with too many correlated predictors as it uses too many parameters which are estimated with a high variance.

In PLS-DA, the response matrix Y is qualitative and is recoded as a dummy block matrix that records the membership of each observation. The PLS regression is then run as if Y was a continuous matrix. Note that this might be wrong from a theoretical point of view, however, it has been previously shown that this works well in practice and many authors have used dummy matrices in PLS for classification [4, 35, 7, 13]. The reader can refer to the article of [4] which gives a formal statistical explanation of the connection between PLS and Linear Discriminant Analysis to explain why the Y -space penalty is not meaningful in this special case.

The PLS-DA model can be formulated as follows:

$$Y = X\beta + e,$$

where β is the matrix of regression coefficients and e the residual matrix. The prediction of a new set of samples is then

$$Y_{new} = X_{new}\beta_{PLS},$$

with $\beta_{PLS} = P(U^T P)^{-1}V^T$, where P is the weight matrix of the X space and U and V are the matrices containing the singular vectors from the X and Y space respectively. The identity of the class membership of each new sample (each row in Y_{new}) can be assigned as the column index of the element with the largest predicted value in this row. This is a naive method for prediction that we call (*maximum* distance). Three other distances are implemented in the `mixOmics`¹ package [29]. The *class* distance allocates the predicted individual x to the class C_k minimizing $dist(x, C_l)$, where $C_k, k = 1, \dots, K$ are the indicator vectors corresponding to each class.

In following Section 3, we illustrate the use of one sparse variants of PLS-DA on a biological data set.

2. PLS for variable selection

From a biological point of view, parsimonious models are needed as the biologists are often interested in the very few relevant genes, proteins or metabolites amongst the thousands for which expression or abundance is measured in high throughput experiments. Their aim is to improve their understanding of the system under study and, if necessary, to perform further validations with reduced experimental costs. From a statistical point of view, parsimonious models are needed in order to be explanatory, interpretable and with a good predictive performance. Many authors have worked on the problem of variable selection with PLS. It first began in the field of chemistry before being applied to or further developed for multivariate omics data analysis.

In this section, we illustrate how variable selection with PLS can be used in the different contexts that were presented in previous Section 1:

- to select predictive variables in multiple linear regression (PLS1, PLS2),
- to select relevant variables while modelling bi-directional relationships between the two data sets (PLS-canonical mode),

¹ <http://www.math.univ-toulouse.fr/~biostat/mixOmics>

- to select discriminative variables in a classification framework (PLS-DA).

We review some of the PLS variants developed for variable selection in the different contexts cited above. Remember that PLS1 considers a single vector of dependent variable Y , whereas PLS2 considers a whole matrix Y of dependent variables.

According to [18], there exists three types of variable selection with PLS:

- *subset selection*: subsets of variables are selected according to a model's performance that is not necessary in line with PLS. PLS is then performed after the variable selection step.
- *dimension-wise selection*: the PLS model is progressively built by removing non informative variables or by adding relevant variables.
- *model-wise elimination*: the PLS model is built with all variables and an internal criteria is used to select the most informative variables.

We will particularly focus on two specific PLS2 variants (Backward PLS-VIP and sparse PLS) that will be numerically compared in Section 3.

2.1. Variable selection for PLS1

2.1.1. Subset selection

GOLPE. [5] first proposed a factorial design to build a PLS model based on different combinations of variables. The same authors then proposed GOLPE (Generating Optimal Linear PLS Estimations, [6]), a D-optimal design that preselects non-redundant variables. A factorial design procedure is then used to run several PLS analyses with different combinations of these variables. Variables that significantly contribute to the prediction of the model are selected, while the others are discarded.

GA-PLS. [32] proposed a novel approach combining Genetic Algorithms (GA) with PLS for variable selection. GA is one of the methods used to solve optimization problems, in particular to select the most informative variables. The response variable used in the GA algorithm is the cross-validated explained variance. GA is performed on a training set, and once PLS is run, the performance of the subset is evaluated by the root mean square error in the test set. Note that GA is very sensitive to the ratio number of variables/number of samples and is not adequate when $n \ll p$.

Clustering approach. The aim of clustering techniques is to reduce the initial set of variables into a subset of new variables which are able to summarize the entire information initially contained. There exists different types of clustering methods. [19] used a descending approach. Principal components or arithmetic mean can also be chosen to represent the clusters of variables. PLS is then performed on these new variables.

Simple regression. Several simple regressions of Y on each variable from X are first performed. Variables with a significant Student test are then selected, based on the assumption that these variables can better predict the response variable. Therefore, noise is removed from the initial data set [19]. PLS is then performed on this subset of variables. The α risk is usually fixed at 5% but this threshold may vary depending on the number of selected variables.

Backward, forward or stepwise. Backward, forward or stepwise multiple regressions are widely used techniques to keep or select the most significant variables in the model. The selection of the variables is based on the choice of the α risk. Backward selection is a descending approach: at first, all variables are included, they are then removed one by one according to the α risk. Forward selection is an ascending approach whereas stepwise selection is a mixture of both. PLS is then performed on the reduced set of variables [19].

2.1.2. Dimension-wise selection

Twenty methods of variable selection were compared in [19] in the context of PLS1 regression, amongst which two may be classified as dimension-wise selection.

Backward Q_{cum}^2 . This approach is a backward selection approach where the variables with the smallest PLS regression coefficient (in absolute value) are removed at each step. Finally, the optimal number of variables to select is defined by the Q_{cum}^2 , a predictive criterion obtained by cross-validation (see [43] for further details).

Backward SDEP. This approach is similar to the once previously described, except that the Q_{cum}^2 criterion is replaced by the square root of the mean square error estimated on a test set (Standard Deviation of Error Prediction).

2.1.3. Model-wise elimination

UVE-PLS. Uninformative Variable Elimination for PLS [10] consists in evaluating the relevancy of each variable in the model through a variable selection criterion, such as the stability of each variable. The uninformative variables are then eliminated. UVE-PLS has been widely applied in analytical chemistry.

IPW-PLS. Iterative Predictor Weighting-PLS [18] multiplies the variables by their importance in the cyclic iterations of the PLS algorithm. It is thus crucial to get a correct PLS model for the purpose of variable selection.

Amongst the 20 methods compared by [19], four can be classified as model-wise elimination:

Correlation method. PLS is first performed with all variables and the PLS dimension is chosen by cross-validation. The correlations between all the PLS latent components and all the variables are then calculated. Variables with at least one non significant correlation coefficient with the latent components are then removed (α risk usually fixed at 5%).

Coefficient method. The adopted strategy is similar to the correlation method, except that it is the ratios between the maximum coefficient of the PLS regression coefficients and the coefficient of each variable that are calculated. Variables with a ratio greater than a threshold fixed by the user are then removed.

Confidence Interval method. This approach is similar to the correlation or coefficient methods. Variables for which their PLS regression coefficient confidence interval includes zero are removed (confidence level usually fixed at 95%). The standard deviation estimator used for the confidence interval is defined in [43].

Jack method. This is the same approach as the confidence interval method, except that the standard deviation for the confidence interval is calculated with Jackknife re-sampling.

2.2. Variable selection for PLS2

2.2.1. Subset selection

In our survey and in the case of PLS2 we did not identify any subset variable selection approach.

2.2.2. Dimension-wise selection

PLS-forward. The PLS-forward consists in selecting variables from an algorithm developed by [26] with a forward approach. The criterion to include a variable within the model is the redundancy index that was introduced by [41]:

$$RI(Y, X) = \frac{\sum_{i=1}^q S_{Y(i)}^2 R_{Y(i)X}^2}{\sum_{i=1}^q S_{Y(i)}^2}, \quad (2)$$

where q is the number of variables in Y , $R_{Y(i)X}^2$ is the squared sample multiple correlation coefficient between the i^{th} variable of Y and the data set X and $S_{Y(i)}^2$, the sample variance of the i^{th} variable of Y . In the PLS regression framework, Y is replaced by the PLS latent components matrix.

IVS-PLS. [33] developed an Interactive Variable Selection approach for PLS. The algorithm is based on the loading vectors of the X variables obtained from the PLS model. Variables with a loading value lower than a threshold fixed by the user are removed from the model. The remaining loading values are then adjusted to keep the unit norm of the loading vector. This step is repeated until there remains only one variable. The best model is then chosen according to a predictive criterion obtained by cross-validation.

Backward PLS-VIP. Stepwise backward and forward regression methods are widely used for variable selection. However, when dealing with the case of omics data characterized by a high multicollinearity, these methods have a poor performance but PLS2 can circumvent this issue. Co-jointly used with the Variable importance in projection (VIP) score [51], backward PLS enables to perform variable selection. The VIP score is calculated for each variable as defined by [43]:

$$VIP_{Hj} = \sqrt{\frac{p}{RI(Y, \xi_1, \xi_2, \dots, \xi_H)} \sum_{l=1}^H RI(Y, \xi_l) w_{lj}^2}, \quad (3)$$

where H is the PLS dimension, (ξ_1, \dots, ξ_H) are the H latent components, p the number of variables in X , $j = 1, \dots, p$ and RI is the redundancy index defined previously.

Detailed backward PLS VIP algorithm.

Start: define the maximum number of variables p_s to be selected (usually arbitrarily chosen by the user).

1. Using cross-validation, determine the PLS dimension H .
2. Perform a PLS regression of Y on X with all the available variables on H dimensions.
3. Remove the variable with the smallest VIP.

Re-iterate these three steps until the number of selected variables is greater than p_s .

Review. [11] have demonstrated the good performance of the VIP selection method compared to other criteria. They also studied the VIP cut-off value to assess variable relevancy. The generally used cut-off value is set to one, but depending on the data properties such as the correlation structure, the authors demonstrated that this cut-off value should actually be greater than one. [27] have also compared the VIP approach for variable selection in the case of PLS1.

2.2.3. Model-wise elimination

PLS-bootstrap. The PLS-bootstrap [27] assumes a multivariate normal distribution for (Y, X) . This method consists in sampling (Y, X) with replacement and in building the PLS model for each sample. From this bootstrap re-sampling, confidence intervals are then estimated for each PLS regression coefficient. A variable is removed if zero is included in the confidence interval.

PLS-VIP. The VIP (Variable Importance in the Projection) method ([51], see description above) was implemented in the SIMCA-P software [46]. The VIP estimates the explanatory performance of the variables within PLS. Variables with $(VIP > 1)$ are then selected.

Sparse PLS. The sparse PLS (sPLS) proposed by [31, 30] was proposed to identify subsets of correlated variables from two different types, e.g. transcriptomics and metabolomics measured on the same samples. It consists in soft-thresholding penalizations of the loading vectors of Y and X to perform variable selection.

The approach is based on Singular Value Decomposition (SVD) of the cross product $M_h = X_h^T Y_h$ that can also be used to solve PLS in a more computationally efficient way. We denote u_h (v_h) the left (right) singular vector from the SVD, for iteration h , $h = 1 \dots H$ where H is the number of performed deflations. These singular vectors are the *loading vectors* in the PLS context. Sparse loading vectors are then obtained by applying l_1 penalization on both u_h and v_h . Therefore, many elements in these vectors are exactly set to zero. The objective function can be written as:

$$\max_{u_h, v_h} \text{cov}(X u_h, Y v_h) \quad (4)$$

$$\text{subject to } \|u_h\| = 1, \|v_h\| = 1 \text{ and } P_{\lambda_1}(u_h) \leq \lambda_1, P_{\lambda_2}(v_h) \leq \lambda_2,$$

where P_{λ_1} and P_{λ_2} are soft-thresholding penalty functions that approximate Lasso penalty functions ($h = 1 \dots H$). The objective function is actually solved by formulating sPLS as a least squares problem using SVD [40, 31]. sPLS minimizes the Frobenius norm between the current cross product matrix and the loading vectors:

$$\min_{u_h, v_h} \|M_h - u_h v_h'\|_F^2 + P_{\lambda_1}(u_h) + P_{\lambda_2}(v_h), \quad (5)$$

where $P_{\lambda_1}(u_h) = \text{sign}(u_h)(|u_h| - \lambda_1)_+$, and $P_{\lambda_2}(v_h) = \text{sign}(v_h)(|v_h| - \lambda_2)_+$ are applied component-wise [40]. They are simultaneously applied on both loading vectors. The problem (5) is solved with an iterative algorithm and the X_h and Y_h matrices are subsequently deflated for each iteration h for either a regression or canonical deflation mode (see [31] for more details).

The penalization parameters can be simultaneously chosen by computing the prediction error with cross-validation. In a regression analysis context however, it is easier to use a criterion such as prediction error Q^2 [43, 31] to help tuning the number of variables. We further discuss this issue in Section 4. In the `mixOmics` R package where the sPLS is implemented, for practical purposes, the user chooses the number of variables to select on each dimension rather than tuning the penalization parameters λ_1 and λ_2 .

Other variant. [12] also developed a sparse PLS version for regression with Lasso penalization, but their approach only permits variable selection on the X data set.

2.3. Variable selection for PLS-canonical mode

Similar to the sPLS approach described above, sparse approaches have been proposed by [47, 36, 50, 30] for a sparse Canonical Correlation Analysis (CCA, [25]) based on the PLS-canonical mode algorithm (see Section 1.3). These methods either include l_1 (Lasso) or l_1 and l_2 (Elastic Net, [54]) penalizations.

PCCA. [47] proposed an approximation of the Elastic Net penalization applied on the loading vectors. This penalization combines the advantages of the ridge regression to obtain a grouping effect and the Lasso for built-in variable selection. Their penalized CCA (PCCA) was applied on brain tumour data sets with gene expression and copy numbers. Later on, the same authors proposed to extend their approach for longitudinal data in a two step procedure involving mixed models and penalized CCA. They illustrated their approach on Single Nucleotide Polymorphisms (SNPs) and repeatedly measured intermediate risk factors [48].

SCCA. [36] applied soft-thresholding penalization using a Lagrange form of the constraints on the loading vectors. They also proposed an extension of their sparse CCA by including adaptive Lasso [53] that includes additional weights in the Lasso constraint. The approach was applied on gene expression and SNPs human data.

sPLS-canonical mode. Similar to [36], [30] implemented sparse PLS with a canonical deflation mode (as presented in Section 1.3) with Lasso penalization as presented above. They compared their approach to [47] and Co Inertia analysis [15] on NCI gene expression data sets measured on two different platforms (cDNA and Affymetrix chips) to highlight complementary information from both platforms. Co-Inertia was found to select redundant information compared to the two other approaches.

sparse CCA. [50] proposed to apply Lasso or fused Lasso [45] in a bound form of the penalties. They extended their sparse CCA to sparse supervised as well as multiple CCA and illustrated their approaches on copy numbers data from diffuse large B-cell lymphoma study.

[47, 36, 50] proposed to tune the number of variables to select by estimating canonical correlation using cross-validation. However, in this particular canonical mode case, the selection of the optimal number of variables remains an open question as the more numerous the variables used to compute the correlation, the larger the correlation coefficient. There must therefore be a trade-off between maximum correlation and the sparsity of the variables.

2.4. Variable selection for PLS-Discriminant Analysis

sPLS-DA. The extension of sparse PLS to a supervised classification framework is straightforward. The response matrix Y of size $(n \times K)$ is coded with dummy variables to indicate the class membership of each sample.

In this specific case, we will only perform variable selection on the X data set, i.e., we want to select the discriminative features that can help predicting the classes of the samples. Therefore, we set $M_h = X_h^T Y_h$ and the optimization problem of the sPLS-DA can be written as:

$$\min_{u_h, v_h} \|M_h - u_h v_h'\|_F^2 + P_\lambda(u_h),$$

with the same notation as in sPLS.

SPLSDA. [13] recently proposed a similar approach except that the variable selection and the classification steps are performed separately - whereas the prediction step in sPLS-DA is directly obtained from the by-products of the sPLS. The authors therefore proposed to apply different classifiers once the variable selection was performed: Linear Discriminant Analysis (SPLSDA-LDA) or a logistic regression (SPLSDA-LOG). The authors also proposed a one-stage approach SGPLS by incorporating sPLS into a generalized linear model framework for a better sensitivity for multiclass classification. These approaches are implemented in the R package `sp1s`. A thorough comparison between the different variants of sPLS-DA can be found in [28], who showed that only SPLSDA-LDA could give similar performance to sPLS-DA, while SGPLS was seriously limited by too large data sets.

In the following Section 3.2, we compare backward PLS-VIP and sPLS-DA on a real biological data set and assess their generalization performance with the maximum and the class distances.

3. Illustration on liver toxicity study

Data. In the liver toxicity study [23], 64 male rats of the inbred strain Fisher 344 were exposed to non-toxic (50 or 150 mg/kg), moderately toxic (1500 mg/kg) or severely toxic (2000 mg/kg) doses of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed at 6, 18, 24 and 48 hours after exposure and the mRNA from liver was extracted. Ten clinical chemistry measurements of variables containing markers for liver injury are available for each object and the serum enzymes levels can be measured numerically. The expression data are arranged in a matrix X of $n = 64$ objects and $p = 3116$ expression levels after normalization and pre-processing, the clinical measurements (Y , $q = 10$) can be predicted using the gene expression matrix in a PLS framework.

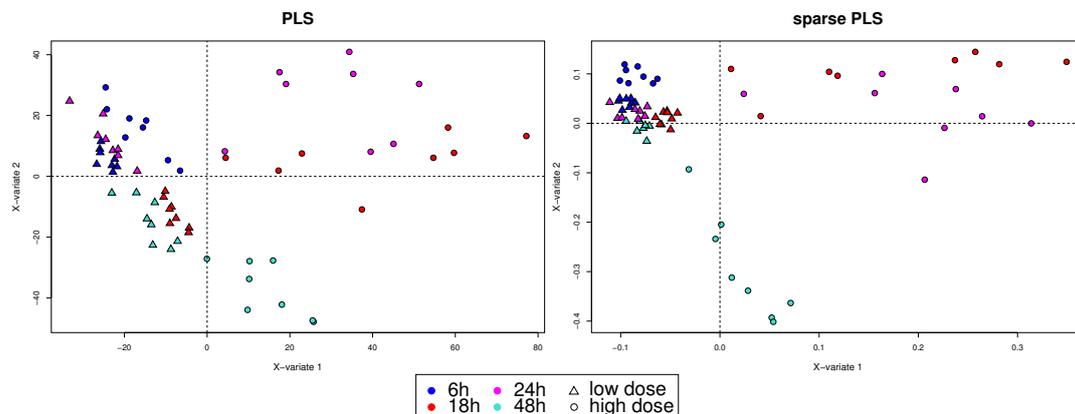


FIGURE 2. Liver toxicity study. Sample representation using the first 2 latent variables from PLS (no variable selection) and sPLS (50 genes selected on each dimension) with *mixOmics*.

3.1. PLS2 regression: examples of graphical outputs using sPLS

The great advantage of PLS and similar projection methods is that they can provide powerful views of the data that are compressed in two to three dimensions. An inspection of the latent variables of loading vectors plots may reveal groups in the data that were previously unknown or uncertain. In this Subsection, we present some of the graphical outputs that can be obtained on the liver toxicity study in using sPLS in a regression framework.

In Figure 2, we compared the sample representations of PLS (with no variable selection) and sparse PLS where 50 genes were selected on each dimension. We can see that variable selection enables better clusters of the samples as only the relevant variables are kept and are used to compute the latent variables. sPLS is therefore able to highlight similarities between the rats which were exposed to either low or high doses of acetaminophen. We can also observe strong differences between the different times of necropsies. The reader can refer to [31] for insightful graphical outputs on transcriptomics and metabolomics yeast data sets.

Correlation circles can be used to represent the variables and to understand how they contribute to the separation of each dimension and as well as illustrating the relative importance of each variable (Fig. 3). In the case of data integration, these valuable graphical outputs give more insight into the correlation structure between the two types of variables (here the selected clinical measurements and the transcripts). The reader can refer to [43, 39] and [20, 29] for an illustration in the context of omics data integration. In particular, [30] showed that the clusters of genes obtained on such correlation circles were related to particular types of tumours.

Recently, further improvements have been done in *mixOmics* to evaluate pair-wise associations between the variables and represent the correlations between the two types of variables using relevance networks (Fig. 4, see also [29]). These inferred networks model both direct and undirected interactions between the two types of variables. They have been shown to bring relevant results as they seem to reproduce known biological pathways [21]. These types of graphical outputs will

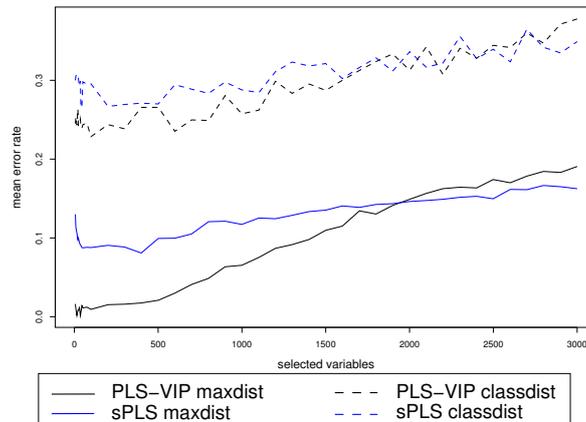


FIGURE 5. Liver toxicity study. Estimated classification error rates in a discriminant analysis framework (10-fold cross-validation averaged 50 times) with respect to the number of selected genes, black (blue) line represent the generalization performance of PLS-VIP (sPLS-DA) using maximum or class distance for prediction.

undoubtedly help the user to give more insight into the data.

3.2. PLS - Discriminant Analysis: numerical comparison of sPLS-DA and backward PLS-VIP

For a PLS-DA framework, we set $Y =$ necropsy time as the class response vector and $X =$ gene expression data and compared the results of backward PLS-VIP and sPLS-DA.

Both approaches generate the same type of outputs. In the case of sPLS-DA, we obtain the sparse loading vectors (u_1, \dots, u_H) which indicate which variables (genes) were selected on each dimension and the H latent components (ξ_1, \dots, ξ_H) . The user can choose the number of variables to select. In the case of PLS-VIP, we obtain the names of the variables that were kept in the backward approach. During the evaluation process, we trained both approaches on training data sets, extracted the names of the selected genes, and tested the prediction of the associated test samples on this same set of genes. We performed 10-fold cross-validation averaged 50 times and computed the classification error rate while varying the variable selection size.

Classification performance. We compared the generalization performances of backward PLS-VIP and sPLS-DA with the maximum and class distances. There is a large difference between the two distances and the maximum distance seems to give the best prediction of the test samples for this multiclass problem. Both variable selection approaches seem to perform similarly, although the backward PLS-VIP has a higher error rate variability than sPLS-DA.

The estimation of the generalization performance also enables to select the ‘optimal’ number of variables to select (the number of variables for which the classification error rate is at its lowest). However, the reader should keep in mind that in such complex and highly dimensional problems, this is a rather challenging question to be addressed.

It is interesting to notice that in overall, both approaches have a similar generalization performance, even though the proportion of commonly selected variables is pretty low: it varied from

30% of overlap for 6-15 selected variables up to 70% overlap for 1,000 selected variables, see Supplemental File. The next important step would therefore be to assess the biological relevancy of these different variable selections with respect to the biological study.

Based on these results, we would advise to use sPLS-DA rather than backward PLS-VIP. In addition, the backward selection is much more computationally demanding than sPLS as PLS-VIP needs to be performed in a stepwise manner for each possible variable selection size. As a result, it took PLS-VIP 1 hour to train instead of few seconds for sPLS-DA for a chosen selection size of 50 variables². Note that the computational time of PLS-VIP could certainly decrease for a larger variable selection size and with a much improved programming code.

More comparisons of sPLS-DA with similar PLS-based approaches can be found on [28]. In this article, sPLS-DA was extensively compared with other sparse Linear Discriminant Analysis approaches (sLDA, [1]) and 3 versions of SPLSDA from [13], as well as some widely used wrapper approaches for gene selection. In many cases, sPLS-DA was found to be clearly competitive to the tested approaches, as well as computationally efficient. Furthermore, the variety of graphical outputs that are proposed in *mixOmics* offer a clear advantage to the other sparse exploratory approaches.

4. Discussion on the validation of the results

Numerical validation. We illustrated the use of PLS for variable selection in a regression/predictive framework. A rather straightforward way to validate the results would be to assess the predictive ability of the obtained models. However, when dealing with omics data, one has to deal with a very small number of samples. Most often, it is impossible to validate the PLS model on an independent data set. An alternative way is to perform cross-validation on the training data set that was used for modelling. This has been used extensively with microarray data analysis, where the number of samples is often ‘large’ (i.e. 50 – 100). However, gathering omics data on matched samples is much more costly and this can lead to extremely small data sets in most cases ($n < 50$). Cross-validation, leave-one-out validation, resampling techniques will allow to compute criteria such as the proportion of explained variance, or the proportion of predicted variance (Q^2). Recently, stability analysis was proposed by [34, 3] to assess the stability of the variable selection (see also [28]). However, the user must keep in mind the limitation of such validation techniques in this small n large p problems.

Biological validation. The use of graphical outputs such as the ones illustrated in Section 3.1 can guide the interpretation of the results. Most importantly, combined with a thorough biological interpretation of the selected transcripts, metabolites, these outputs will give a clear indication whether the proposed model answers the biological questions. The use of biological softwares (GeneGo [2], Ingenuity Pathways Analysis³, to cite a few) or a thorough search in the biological literature to further investigate if these selected variables have a biological meaning with respect to the study is the ultimate way to validate the results. The statistician analysing such data must keep in mind the biological question to be answered.

² run on a 2.66GHz machine with 4GB of RAM using R

³ Ingenuity® Systems, www.ingenuity.com

How many variables to select? Another critical issue is the optimal number of variables to select. In a regression/classification framework, this can be answered using cross-validation and different criteria such as Q_{cum}^2 or the classification error rate. In practice however, this may not be interesting for the biologist. The selection size might be too small and in that case the results cannot be processed further through biological software (not enough information), or, conversely, the selection size might be too large which makes an experimental validation impossible. Therefore, it may often happen that the number of variables to be selected has to be guided by the biologist rather than by using statistical criteria.

Conclusion

PLS-based methods are useful and versatile approaches for modelling, monitoring and predicting complex problems and data structures encountered within the omics field. The other virtue of such approaches is that their results can be graphically displayed in many different ways. In many studies, PLS-based methods were shown to bring biologically relevant results as they are able to capture the dominant, latent properties of the studied system. The use of PLS and derived methods for data reduction is becoming increasingly relevant to handle the current explosion of the size of analytical data sets obtained from any biological system.

In this review, we presented the recent developments in PLS modelling for variable selection and demonstrated the usefulness of such improvements in PLS to deal with the new challenges posed by the systems biology arena. Variable selection within PLS can select relevant information while integrating omics data sets. The graphical outputs inherent from PLS are a valuable addition to enable a clear visualization of the results, as illustrated on one data set. In a discriminant analysis framework and on a real data set, we compared the classification performance of two PLS-based variable selection approaches: backward PLS-VIP and sPLS-DA. sPLS-DA was found to be the most efficient in terms of generalization ability and computational performance. This type of approach is easily applicable to systems biology studies and will undoubtedly help in addressing fundamental biological questions and in understanding systems as a whole.

Acknowledgment

We would like to thank the two reviewers whose comments helped improve the clarity of the manuscript.

References

- [1] M. Ahdesmäki and K. Strimmer. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Stat.*, 2009.
- [2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [3] F. Bach. Model-consistent sparse estimation through the bootstrap. Technical report, Laboratoire d'informatique de l'Ecole Normale Supérieure, Paris, 2009.
- [4] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.
- [5] M. Baroni, S. Clementi, G. Cruciani, G. Costantino, D. Rignanelli, and E. Oberrauch. Predictive ability of regression models: Part II. Selection of the best predictive PLS model. *Journal of chemometrics*, 6(6):347–356, 1992.

- [6] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quantitative Structure-Activity Relationships*, 12(1):9–20, 1993.
- [7] A.L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32, 2007.
- [8] N.A. Butler and M.C. Denham. The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society B*, 62(3):585–594, 2000.
- [9] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, and J. Trygg. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52(6):1181–1191, 2007.
- [10] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, and C. Sterna. Elimination of uninformative variables for multivariate calibration. *Anal. Chem*, 68(21):3851–3858, 1996.
- [11] IG. Chong and CH. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1-2):103–112, 2005.
- [12] H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.
- [13] D. Chung and S. Keles. Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology*, 9(1):17, 2010.
- [14] S. de Jong. Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993.
- [15] S. Dolédec and D. Chessel. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater Biology*, 31(3):277–294, 1994.
- [16] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, and S. Wold. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and bioanalytical chemistry*, 380(3):419–429, 2004.
- [17] J.M. Fonville, S.E. Richards, R.H. Barton, C.L. Boulange, T. Ebbels, J.K. Nicholson, E. Holmes, and M.E. Dumas. The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics*, 2010.
- [18] M. Forina, C. Casolino, and C.P. Millan. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Journal of Chemometrics*, 13(2):165–184, 1999.
- [19] JP. Gauchi and P. Chagnon. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, 58(2):349–363, 2001.
- [20] I González, S Déjean, P Martin, O Gonçalves, P Besse, and A Baccini. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(2):173–199, 2009.
- [21] I. González, K-A. Lê Cao, M. Davis, and S. Déjean. Insightful graphical outputs to explore relationships between two ‘omics’ data sets. Technical report, Université de Toulouse, 2011.
- [22] C. Goutis. Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics*, 24(2):816–824, 1996.
- [23] A.N. Heinloth, R.D. Irwin, G.A. Boorman, P. Nettesheim, R.D. Fannin, S.O. Sieber, M.L. Snell, C.J. Tucker, L. Li, G.S. Travlos, et al. Gene Expression Profiling of Rat Livers Reveals Indicators of Potential Adverse Effects. *Toxicological Sciences*, 80(1):193–202, 2004.
- [24] I.S. Helland. Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58(2):97–107, 2001.
- [25] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.
- [26] A. Lazraq and Cléroux. The PLS multivariate regression model: testing the significance of successive PLS components. *Journal of chemometrics*, 15(6):523–536, 2001.
- [27] A. Lazraq, R. Cléroux, and JP. Gauchi. Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, 66(2):117–126, 2003.
- [28] K-A Lê Cao, S. Boitard, and P. Besse. Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. Technical report, University of Queensland, 2011.

- [29] K-A. Lê Cao, I. González, and S. Déjean. integrOmics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21):2855–2856, 2009.
- [30] K-A. Lê Cao, P.G.P. Martin, C. Robert-Granié, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(34), 2009.
- [31] K-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. Sparse PLS: Variable Selection when Integrating Omics data. *Statistical Application and Molecular Biology*, 7((1):37), 2008.
- [32] R. Leardi, R. Boggia, and M. Terrile. Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, 6(5):267–281, 1992.
- [33] F. Lindgren, P. Geladi, S. Rännar, and S. Wold. Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. *Journal of Chemometrics*, 8(5):349–363, 1994.
- [34] N. Meinshausen and P. Bühlmann. Stability selection. Technical report, ETH Zurich, 2008.
- [35] D.V. Nguyen and D.M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39, 2002.
- [36] E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.
- [37] A. Phatak, PM Reilly, and A. Penlidis. The asymptotic variance of the univariate PLS estimator. *Linear Algebra and its Applications*, 354(1-3):245–253, 2002.
- [38] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*, pages 34–51, 2006.
- [39] G Saporta. *Probabilités analyse des données et statistique*. Technip, 2006.
- [40] Haipeng Shen and Jianhua Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99:1015–1034, 2008.
- [41] D. Stewart and W. Love. A general canonical index. *Psychology Bulletin*, 70(3):160–163, 1968.
- [42] Y. Tan, L. Shi, W. Tong, GT Gene Hwang, and C. Wang. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Computational Biology and Chemistry*, 28(3):235–243, 2004.
- [43] M. Tenenhaus. *La régression PLS: théorie et pratique*. Editions Technip, 1998.
- [44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [45] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.
- [46] A. Umetri. SIMCA-P for windows, Graphical Software for Multivariate Process Modeling. Umea, Sweden, 1996.
- [47] S. Waaijenborg, V. de Witt Hamer, C. Philip, and A.H. Zwinderman. Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(3), 2008.
- [48] S. Waaijenborg and A.H. Zwinderman. Association of repeatedly measured intermediate risk factors for complex diseases with high dimensional SNP data. *Algorithms for Molecular Biology*, 5(1):17, 2010.
- [49] J.A. Wegelin et al. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical report, Seattle: Department of Statistics, University of Washington, 2000.
- [50] D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515, 2009.
- [51] S. Wold, E. Johansson, and M. Cocchi. *3D QSAR in Drug Design; Theory, Methods, and Applications, PART III ESCOM*. KLUWER/ESCOM, 1993.
- [52] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.
- [53] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.
- [54] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.