View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Queensland eSpace

Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 35

A Sparse PLS for Variable Selection when Integrating Omics Data

Kim-Anh Lê Cao, *INRA UR 631 and Université de Toulouse*
Debra Rossouw, *University of Stellenbosch*
Christèle Robert-Granié, *INRA UR 631*
Philippe Besse, *Université de Toulouse*

Recommended Citation:
Lê Cao, Kim-Anh; Rossouw, Debra; Robert-Granié, Christèle; and Besse, Philippe (2008) "A Sparse PLS for Variable Selection when Integrating Omics Data," *Statistical Applications in Genetics and Molecular Biology*: Vol. 7: Iss. 1, Article 35.
DOI: 10.2202/1544-6115.1390

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

A Sparse PLS for Variable Selection when Integrating Omics Data

Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié, and Philippe Besse

Abstract

Recent biotechnology advances allow for multiple types of omics data, such as transcriptomic, proteomic or metabolomic data sets to be integrated. The problem of feature selection has been addressed several times in the context of classification, but needs to be handled in a specific manner when integrating data. In this study, we focus on the integration of two-block data that are measured on the same samples. Our goal is to combine integration and simultaneous variable selection of the two data sets in a one-step procedure using a Partial Least Squares regression (PLS) variant to facilitate the biologists' interpretation. A novel computational methodology called ``sparse PLS'' is introduced for a predictive analysis to deal with these newly arisen problems. The sparsity of our approach is achieved with a Lasso penalization of the PLS loading vectors when computing the Singular Value Decomposition.

Sparse PLS is shown to be effective and biologically meaningful. Comparisons with classical PLS are performed on a simulated data set and on real data sets. On one data set, a thorough biological interpretation of the obtained results is provided. We show that sparse PLS provides a valuable variable selection tool for highly dimensional data sets.

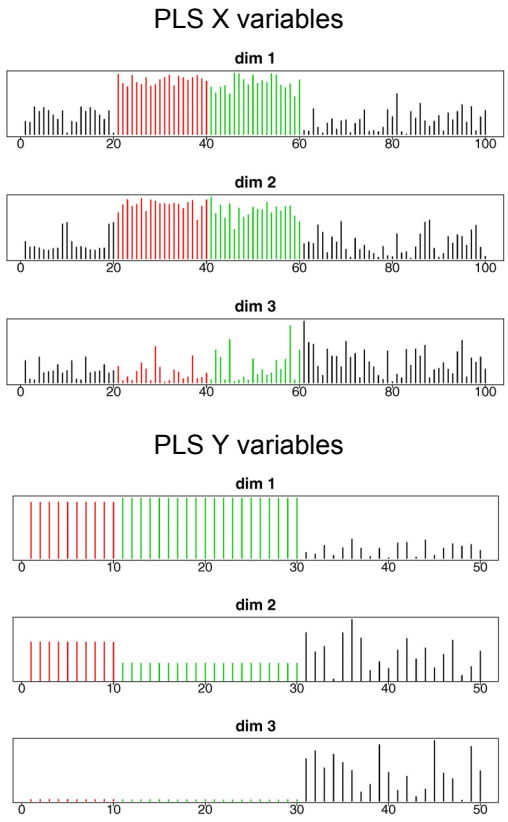
KEYWORDS: joint analysis, two-block data set, multivariate regression, dimension reduction

Author Notes: We are very grateful to Yves Gibon (Max Planck Institute of Molecular Plant Physiology) who kindly provided the full Arabidopsis data set. We would like to thank the anonymous reviewers and the editor for their helpful comments on the manuscript. KALC and CRG are affiliated with the Institut National de la Recherche Agronomique, Station d'Amelioration Genetique Animale. DR is affiliated with the Institute of Wine Biotechnology. KALC (is also) and PB are affiliated with the Institut de Mathematiques de Toulouse, UMR CNRS 5219.

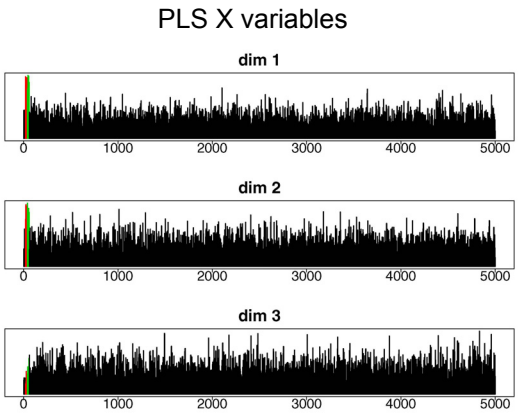
Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Erratum

The PLS X variables and PLS Y variables in Figure 1 (Page 12) should appear as follows:



The PLS X variables in Figure 6 (Page 25) should appear as follows:



Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Introduction

Motivation. Recent technological advances enable the monitoring of an unlimited quantity of data outputs from various sources. These data are generated by different analytical platforms, hence the possibility exists for integration among the different types of data, such as transcriptomic, proteomic or metabolomic data. This integrative biological approach enables an improved understanding of some underlying biological mechanisms and interactions between functional levels to be obtained. This is conditional upon the successful incorporation of several omics data-types that are characterized by many variables but not necessarily many samples or observations. In this high dimensional setting, it is absolutely crucial to select genes, proteins or metabolites in order to overcome computational limits (from a mathematical and statistical point of view) and to facilitate biological interpretation. Therefore, our quest for sparsity is motivated by the biologists' needs for the separation of the useful information related to the study from irrelevant information caused by experiment inaccuracies. The resulting variable selection might also enable a feasible/targeted biological validation at a reduced experimental cost. In this paper, we especially focus on the integrative context, which is the main goal of omics data. For example, a biological study might aim to explain the q metabolites by the p transcripts that are measured on the same n samples. In this typical case, $n << p + q$.

In this study, we propose a sparse version of the Partial Least Squares regression (PLS, Wold 1966), that aims to combine variable selection *and* modelling in a one-step procedure for such a problem. Our sparse PLS is based on Lasso penalization (Tibshirani, 1996) and is obtained by penalizing a sparse Singular Value Decomposition (SVD), as proposed by Shen and Huang 2008, by using a PLS variant with SVD (Lorber et al., 1987). This approach deals with integration or joint analysis problems, which cannot be solved with usual feature selection approaches proposed in classification or discrimination studies, where there is only one data set to analyze. Hence, multiple testing to identify differentially expressed genes does not apply here, as well as other classification methods that have been applied to transcriptomic data sets. In the latter case, many authors (among them: Guyon et al. 2002; Lê Cao et al. 2007) have applied feature selection methods to microarray data. These methods have proved to select biologically meaningful genes lists. However, in the present context, the feature selection aim should be combined with modelling two-block data sets. Very few approaches have been proposed to deal with these newly arisen problems, especially for *predicting* one group of variables from the other group. Several approaches that seek linear combinations of both

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

groups of variables can answer this biological problem. However, they are often limited by collinearity or by ill posed problems that require regularization techniques, such as l_1 (Lasso) or l_2 (Ridge) penalizations.

Background and related work. Partial Least Squares regression (PLS, Wold 1966) is a well known regression technique that was initially applied in chemometrics. When faced with collinear matrices the stability of PLS gives it a clear advantage over CCA, multiple linear regression, ridge regression or other regression techniques. Furthermore, since Wold's original approach, many variants have arisen (SIMPLS, de Jong 1993, PLS1 and 2, PLS-A, PLS-SVD, see Wegelin 2000 for a survey) that provide the user with a solution for almost any problem. We will describe and discuss some of these variants in this study. PLS is an algorithmic approach that has often been criticized for its lack of theoretical justifications. However, this computational and exploratory approach is extremely popular thanks to its efficiency, and much work still needs to be done to demonstrate all the statistical properties of the PLS (see for example Krämer 2007; Chun and Keles 2007 who recently addressed some theoretical developments of the PLS).

PLS has recently been successfully applied to biological data, such as gene expression (Datta, 2001), integration of gene expression and clinical data (with bridge PLS, Gidskehaug et al. 2007), integration of gene expression and ChIP connectivity data (Boulesteix and Strimmer, 2005) and more recently for reconstructing interaction networks from microarray data (Pihur et al., 2008). We can also mention the study of Culhane et al. (2003) who applied Co-Inertia Analysis (CIA, Dolédec and Chessel 1994) from which PLS is a particular case, in a cross platform comparison of microarray data.

In the context of feature selection from both data sets, one closely related work that has proved to give biologically meaningful results is the O2PLS model (Trygg and Wold, 2003). Variable selection was added to this model by Bylesjö et al. (2007) for combining and selecting transcript and metabolite data in *Arabidopsis Thaliana* in a regression framework. O2PLS decomposes each data set into three structures (predictive, unique and residual). The most dominant correlation and covariance in both sample directions and variable directions are extracted and can be interpreted. Variable selection is then performed on the correlation loadings with a permutation strategy in a two-step procedure.

More recently, Waaijenborg et al. (2008) and Chun and Keles (2007) both adapted Elastic Net regularization (Zou and Hastie, 2005) in the PLS, either in a canonical framework or in a regression framework. They directly penalized

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

the optimization problem. Both approaches seem promising, as Chun and Keles (2007) demonstrated that the PLS consistency property does not hold when $n << p+q$. However, it would be useful to show the biological relevance of their results. Nevertheless, their studies show the need for developing such integrative methods for biological problems.

Our contribution and results. We propose a sparse PLS approach that combines both integration and simultaneous variable selection on the two data sets in a one-step strategy. We show that our approach is applicable on high-throughput data sets and provides more relevant results compared to PLS.

Outline of the paper. A brief introduction to PLS will be given before describing the sparse PLS method. We describe how to add sparsity into PLS with a Lasso penalization combined with SVD computation (Shen and Huang, 2008). We then assess the validity of the approach on one simulated data set and on three real data sets. We compare and discuss the results with reference to a classical PLS approach. We also provide a full biological interpretation of the results obtained on a typical integrative study of wine yeast which combines relative transcript levels and metabolite concentrations. We show that sparse PLS highlights the most essential transcripts that are relevant to specific metabolites.

1 Methods

1.1 PLS

The PLS regression looks for a decomposition of centered (possibly standardized) data matrices X ($n \times p$) and Y ($n \times q$) in terms of components scores, also called latent variables, $(\xi_1, \xi_2 \dots \xi_H)$ and $(\omega_1, \omega_2 \dots \omega_H)$, that are n-dimensional vectors, and associated loadings, $(u_1, u_2 \dots u_H)$ and $(v_1, v_2 \dots v_H)$, that are p and q- dimensional vectors respectively, to solve the following optimization problem (Burnham et al., 1996):

$$\max_{||u_h||=1, ||v_h||=1} cov(X_{h-1}u_h, Yv_h) \quad (1)$$

where X_{h-1} is the residual (deflated) X matrix for each PLS dimension $h = 1 \dots H$. Note that problem (1) is equivalent to solve: $\max cov(\xi_h, \omega_h)$.

Many PLS variants exist depending on the way the matrix Y is deflated. There exists either a symmetric way ("PLS-mode A") or an asymmetric way

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

("PLS2") (Tenenhaus, 1998; Wegelin, 2000)) to deflate Y and the models consequently differ. In this study we will focus on a regression framework, that is, an asymmetric deflation of Y.

In the case of a regression mode, the models of X- and Y-spaces are respectively (Hoskuldsson, 1988):

$$X = \Xi C^T + \varepsilon_1, \qquad Y = \Xi D^T + \varepsilon_2 = XB + \varepsilon_2, \tag{2}$$

where Ξ $(n \times H)$ is the matrix of PLS components ξ_h and B $(p \times H)$ is the matrix of regression coefficients. The column vectors of C and D are defined as $c_h = X_{h-1}^T \xi_h / (\xi_h' \xi_h)$ and $d_h = Y_{h-1}^T \xi_h / (\xi_h' \xi_h)$. The matrices ε_1 $(n \times p)$ and ε_2 $(n \times q)$ are the residual matrices and $h = 1 \ldots H$.

Other PLS alternatives exist depending on whether X and Y are separately or directly deflated. The latter case uses the cross product $M = X^T Y$ and the SVD decomposition of M. We will discuss these approaches in sections 1.2 and 1.4. Note that in any case, all PLS variants are equivalent when computing the first dimension of the PLS.

1.2 SVD decomposition and PLS-SVD

We recall the SVD decomposition and the principle of the PLS-SVD approach for a better understanding of our sparse PLS approach.

1.2.1 Singular value decomposition

Any real r-rank M $(p \times q)$ matrix can be decomposed into three matrices U, Δ, V as follows:

$$M = U \Delta V^T,$$

where $U(p \times r)$ and $V(q \times r)$ are orthonormal and $\Delta(r \times r)$ is a diagonal matrix whose diagonal elements δ_k $(k = 1 \ldots r)$ are called the singular values. The singular values are equal to the square root eigenvalues of the matrices $M^T M$ and $M M^T$. One interesting property that will be useful in our sparse PLS approach is the fact that the column vectors of U and V, noted $(u_1, \ldots, u_r)$ and $(v_1, \ldots, v_r)$ (resp. called left and right singular vectors) exactly correspond to the PLS loadings of X and Y if $M = X^T Y$.

1.2.2 PLS-SVD

In PLS-SVD, the SVD decomposition of $M = X^T Y$ is performed only once. For each dimension h, M is directly deflated by its rank-one approximation ($M_h = M_{h-1} - \delta_h u_h v_h'$). This computationally attractive approach may

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

however lead to non mutually orthogonal latent variables, in contrast with the properties of PLS2 ($\xi_s'\xi_r = 0, r < s$) or PLS-mode A ($\xi_s'\xi_r = 0$ *and* $\omega_s'\omega_r = 0, r < s$).

1.3 Lasso penalization

Shen and Huang (2008) proposed a sparse PCA approach using the SVD decomposition of $X = U\Delta V^T$ by penalizing the PCA loading vector v for each PCA dimension $h = 1 \dots H$. The optimization problem to solve is

$$\min_{u,v} ||X - uv'||_F^2 + P_\lambda(v), \tag{3}$$

where $||X - uv'||_F^2 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - u_i v_j)^2$ and $P_\lambda(v) = \sum_{j=1}^p p_\lambda(|v_j|)$ is a penalty function.

Solving (3) is performed in an iterative way, as described below:

- Decompose $X = U\Delta V^T$, $X_0 = X$
- For h in 1..H:
  1. Set $v_{old} = \delta_h v_h^\star$, $u_{old} = u_h^\star$, where $v_h^\star$ and $v_h^\star$ are unit vectors
  2. Until convergence of u_{new} and v_{new}:
     (a) $v_{new} = g_\lambda(X_{h-1}^T u_{old})$
     (b) $u_{new} = X^T v_{new} / ||X_{h-1}^T v_{new}||$
     (c) $u_{old} = u_{new}$, $v_{old} = v_{new}$
  3. $v_{new} = v_{new} / ||v_{new}||$
  4. $X_h = X_{h-1} - \delta_h u_{new} v_{new}'$

where $g_\lambda(y) = sign(y)(|y| - \lambda)_+$ is the soft-thresholding function. Other functions were proposed by the authors, such as the hard thresholding function (Donoho and Johnstone, 1994), which was not considered in this study (see section 4). In our particular PLS case, we are interested in penalizing both loading vectors u_h and v_h to perform variable selection in both data sets, $h = 1 \dots H$. Indeed, one interesting property of PLS is the direct interpretability of the loading vectors as a measure of the relative importance of the variables in the model (Wold et al., 2004). Our optimization problem becomes:

$$\min_{u,v} ||M - uv'||_F^2 + P_{\lambda_1}(u) + P_{\lambda_2}(v), \tag{4}$$

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

which is iteratively solved in the above algorithm by replacing X by M and the steps 2.**a.** and 2.**b.** by:

$$v_{new} = g_{\lambda_1}(M_{h-1}^T u_{old})$$

$$u_{new} = g_{\lambda_2}(M_{h-1} v_{old}),$$

where $g_{\lambda_1}(y) = sign(y)(|y| - \lambda_1)_+$ and $g_{\lambda_2}(y) = sign(y)(|y| - \lambda_2)_+$ are the soft-thresholding functions. The sparse PLS algorithm is described in detail in next the section.

1.4 Sparse PLS

It is easy to understand that during the deflation step of the PLS-SVD, $M_h \neq X_h^T Y_h$. This is why we propose to separately compute X_h and Y_h, then to decompose $\tilde{M}_h = X_h^T Y_h$ at each step, and finally to extract the first pair of singular vectors. As Hoskuldsson (1988) explains, taking one pair of loadings (u_h, v_h) at a time will lead to a biggest reduction of the total variation in the X and Y-spaces. In our approach, the SVD decomposition will provide a useful tool for selecting variables from each data set.

We now detail the sparse PLS algorithm (sPLS) based on the iterative PLS algorithm (see Tenenhaus 1998) and SVD computation of $\tilde{M}_h$ for each dimension.

1. $X_0 = X \qquad Y_0 = Y$
2. For h in 1..H:
   (a) Set $\tilde{M}_{h-1} = X_{h-1}^T Y_{h-1}$
   (b) Decompose $\tilde{M}_{h-1}$ and extract the first pair of singular vectors $u_{old} = u_h$ and $v_{old} = v_h$
   (c) Until convergence of u_{new} and v_{new}:
      i. $u_{new} = g_{\lambda_2}(\tilde{M}_{h-1} v_{old})$, norm u_{new}
      ii. $v_{new} = g_{\lambda_1}(\tilde{M}_{h-1}^T u_{old})$, norm v_{new}
      iii. $u_{old} = u_{new}$, $v_{old} = v_{new}$
   (d) $\xi_h = X_{h-1} u_{new} / u'_{new} u_{new}$
   $\omega_h = Y_{h-1} v_{new} / v'_{new} v_{new}$
   (e) $c_h = X_{h-1}^T \xi_h / \xi'_h \xi_h$
   $d_h = Y_{h-1}^T \xi_h / \xi'_h \omega_h$
   $e_h = Y_{h-1}^T \omega_h / \omega'_h \omega_h$

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

(f) $X_h = X_{h-1} - \xi_h c_h'$

(g) $Y_h = Y_{h-1} - \xi_h d_h'$

Note that in the case where there is no sparsity constraint ($\lambda_1 = \lambda_2 = 0$) this approach is reduced to a classical PLS.

1.5 Missing data

When dealing with biological data, it is very common to be confronted with missing data. In order not to lose too much information, an interesting approach for substituting each missing data with a value could be provided by the Non LInear Estimation by Iterative Partial Least Squares (NIPALS, Wold 1966). This method has been at the origin of PLS and allows performing PCA with missing data on each data set. Details of the algorithm can be found in Tenenhaus (1998). Several studies show that the convergence of NIPALS and its good estimation are limited by the number of missing values (20-30% of the whole data set), see for example Dray et al. (2003). NIPALS is now implemented in the `ade4` package (Thioulouse et al., 1997).

1.6 Tuning criteria and evaluation

1.6.1 Lasso penalization

There are two ways of tuning the two penalization parameters λ_1^h and λ_2^h for each PLS dimension, *i.e* choosing the degree of sparsity of each loading vector. The first and rather straightforward way is to use k-fold cross validation or leave-one-out cross validation and compute the prediction error ("*RMSEP*" see section 1.6.3) in order to choose the optimal sparse loading vectors.

When the number of samples is small, the estimated prediction error might be biased and the optimal degree of sparsity can not be computed. In this particular case, it is advisable to arbitrarily choose the number of non zero components in each loading vector u_h, v_h or in both, for each dimension h. This tuning option is adequate for our application on biological data sets, as many omics data are still unknown (*e.g* associated functions, annotations). Variable selections that are limited in size may thus not allow for correct assessment of the results by biologists. In this paper, we therefore focused on the latter approach when analyzing biological data sets. This approach was also proposed by Zou and Hastie (2005) in their `R` package `elasticnet` for their sparse PCA and by Shen and Huang (2008). The first approach is advised when n is large enough (*e.g* $n \geq 100$). However, we used this approach

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

in the simulation study in section 2.1.2.

In our framework, different sparsity degrees can be chosen for both loading vectors and for each dimension.

1.6.2 Choice of the PLS dimension

Marginal contribution of the latent variable ξ_h. In the case of a regression context, Tenenhaus (1998) proposed to compute a criteria called Q_h^2 that measures the marginal contribution of ξ_h to the predictive power of the PLS model by performing cross validation computations. As the number of samples n is usually small in this case, we prefer to use leave-one-out cross validation. Q_h^2 is computed for all Y variables and is defined as

$$Q_h^2 = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}},$$

where $PRESS_h^k = \sum_{i=1}^{n}(y_i^k - \hat{y}_{h(-i)}^k)^2$ is the PRediction Error Sum of Squares and $RSS_h^k = \sum_{i=1}^{n}(y_i^k - \hat{y}_{hi}^k)^2$ is the Residual Sum of Squares for the variable k and the PLS dimension h.

We define the estimated matrix of regression coefficients $\hat{B}$ of B, using the same notation as in equation (2): $\hat{B} = U^* D^T$ where $U^* = U(C^T U)^{-1}$ (see De Jong and Ter Braak 1994; Tenenhaus 1998) and where the column vectors of U are the loading vectors $(u_1, \ldots, u_h)$, $h = 1 \ldots H$. For any sample i, we can predict $\hat{y}_{hi}^k = x_{hi}\hat{B}_{h(-i)}^k$.

This criteria is the one adopted in the *SIMCA-P* software (developed by S. Wold and Umetri 1996). The rule to decide if ξ_h contributes significantly to the prediction is if

$$Q_h^2 \geq (1 - 0.95^2) = 0.0975.$$

However, the choice of the PLS dimension still remains an open question that has been mentioned by several authors (see Mevik and Wehrens 2007; Boulesteix 2004). In our particular biological context, we can show that graphical representations of the samples facilitate this choice as the plots of (ξ_h, ξ_{h+1}) and (ω_h, ω_{h+1}) do not have a biological meaning if h is too large. In fact, our results (see below) show that all relevant information in terms of identification of the measured biological effects can be extracted from 3 dimensions.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

1.6.3 Evaluation

RMSEP For a regression context, Mevik and Wehrens (2007), Boulesteix (2004) in the `R` `pls` and `plsgenomics` packages proposed to compute the Root Mean Squared Error Prediction criterion (RMSEP) with cross validation in order to choose the H parameter. As we already suggested the use of the Q^2_h criterion for this issue, we suggest that the RMSEP criterion is used instead as a way of evaluating the predictive power of each Y variable between the original non penalized PLS and the sPLS in the next section.

Note that the Q^2_h criteria is closely related to the RMSEP ($= \sqrt{PRESS_{kh}}$) and gives a more general insight of the PLS, whereas the RMSEP needs to be computed for each variable k in Y.

2 Validation studies

The evaluation of any statistical approach is usually performed with simulated data sets. In the context of biological data, however, simulation is a difficult exercise as one has to take into account technical effects that are not easily identifiable even on real data sets. We first propose to simulate as realistically as possible two-block data sets in a regression framework, to answer the following questions: does the sparse PLS select relevant variables? Does the variable selection simultaneously performed on both data sets improve the predictive ability of the model, compared to the PLS which includes all variables in the model? Once these questions are answered, the next step is to show that our approach is applicable on biological data sets with various complexities, and that our approach may give potentially relevant results from a statistical point of view compared to PLS. Finally, in the next section, we provide a detailed biological interpretation for one of the data sets, and show that the sparse PLS provides better answers to key biological questions compared to the PLS.

2.1 Simulation study

2.1.1 Simulation design

As proposed by Chun and Keles (2007), this simulation is designed to compare the prediction performance of the PLS and the sPLS in the case where the relevant variables are not governed by a latent variable model. In this setting, we also added two cross conditions to add complexity to this setting. We set $p = 5000$ genes, $q = 50$ response variables and $n = 40$ samples, all with base error model being Gaussian with unit variance. We defined the mean vectors

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

μ_1 and μ_2 as follows and divided the samples into consecutive blocks of 10, denoted by the sets (a, b, c, d), where

$$\mu_{1i} = \begin{cases} -2 & \text{if } i \in a \cup b \\ +2 & \text{otherwise.} \end{cases}$$

$$\mu_{2i} = \begin{cases} -1.5 & \text{if } i \in a \cup c \\ +1.5 & \text{otherwise.} \end{cases}$$

For the first 20 genes, we generated 20 columns of X from a multivariate normal with an AR(1) covariance matrix with auto correlation $\rho = 0.9$. These genes will get a strong Y response, but should not be of interest in the model. The next 40 genes have the mean structure μ_1 or μ_2:

$$\begin{aligned} x_{ij} = &\quad \mu_{1i} + \epsilon_{ij}, \quad j = 21 \ldots 40, \quad i = 1 \ldots n \\ x_{ij} = &\quad \mu_{2i} + \epsilon_{ij}, \quad j = 41 \ldots 60, \quad i = 1 \ldots n. \end{aligned}$$

The next genes have the mean structure U_m and are generated by $X_j = U_m + \epsilon_j$, $m = 1 \ldots 4$, with

$$\begin{aligned} U_1 = &\quad -1.5 + 1.5\mathbb{1}_{u_{ij} \le 0.4}, \quad 1 \le i \le n, \quad 61 \le j \le 80, \\ U_2 = &\quad +1.5 - 1.5\mathbb{1}_{u_{ij} \le 0.7}, \quad 1 \le i \le n, \quad 81 \le j \le 100, \\ U_3 = &\quad -2 + 2\mathbb{1}_{u_{ij} \le 0.3}, \quad 1 \le i \le n, \quad 101 \le j \le 120, \\ U_4 = &\quad +2 - 2\mathbb{1}_{u_{ij} \le 0.3}, \quad 1 \le i \le n, \quad 121 \le j \le 140, \end{aligned}$$

where $u_{ij} \sim \mathcal{U}(0, 1)$ and ϵ_j are i.i.d random vectors from $\mathcal{N}(0, \mathbb{1}_n)$. In all cases, $\epsilon_{ij} \sim \mathcal{N}(0, 1)$, which is also how the remaining 4860 genes are defined.

The response variables Y_{ik} follow $Y_k = X\beta_1 + e_k$, $k = 1 \ldots 10$, with

$$\beta_{1j} = \begin{cases} 10 & \text{if } 1 \le j \le 20, \\ 8 & \text{if } 21 \le j \le 40, \\ 4 & \text{if } 21 \le j \le p, \end{cases}$$

and $Y_k = X\beta_2 + e_k$, $k = 11 \ldots 20$ with

$$\beta_{2j} = \begin{cases} 10 & \text{if } 1 \le j \le 20, \\ 4 & \text{if } 21 \le j \le 40, \\ 8 & \text{if } 21 \le j \le p, \end{cases}$$

and $Y_k \sim e_k$ for $k = 21 \ldots 50$ with $e_k \sim \mathcal{N}(0, \mathbb{1}_n)$.

In this simulation setting, the tested methods should highlight the genes X_j, $j = 11 \ldots 40$ and the response variables Y_k, $k = 1 \ldots 30$, which are related either to a μ_1 or a μ_2 effect.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Table 1: Average RMSEP (standard error) for each PLS dimension for 50 simulated data sets, with leave-one-out cross validation. For the sPLS, 30 response variables were selected on each dimension.

	PLS	sPLS 20 genes	sPLS 40 genes	sPLS 60 genes	sPLS 100 genes
dim 1	0.926 (0.009)	0.839 (0.046)	0.731 (0.038)	0.677 (0.030)	0.671 (0.019)
dim 2	0.922 (0.009)	0.558 (0.021)	0.564 (0.020)	0.585 (0.027)	0.611 (0.027)
dim 3	0.921 (0.008)	0.557 (0.002)	0.566 (0.022)	0.582 (0.0243)	0.579 (0.071)

2.1.2 Prediction performance

X and Y were simulated 50 times and we performed leave-one-out cross validation on each data set. For sparse PLS, we arbitrarily chose to select 30 response variables for each dimension h, $h = 1 \dots 3$ and let the number of selected genes vary. For PLS, no penalization was applied, so that all Y variables were modelled with respect to the whole X data set for each simulation run.

The RMSEP for each response variable, each test set and each dimension was computed and averaged in Table 1. These results show that sPLS improves the predictive ability of the model compared to the PLS. After dimension $H = 2$, neither sPLS nor PLS get a significant decrease in the average RMSEP. This is in agreement with our simulation design, in which only two latent effects, μ_1 and μ_2, are included. The next section shows that these effects are indeed highlighted by PLS and sPLS in the first 2 dimensions. Furthermore, if one had to choose the optimal number of genes to select, the best solution would be to select between 20 and 40 genes on the first 2 dimensions, as the lowest average RMSEP is obtained for dimension 2. This is in agreement with the simulation design.

2.1.3 Variable selection

In this part, we compare the loading vectors (u_1, u_2, u_3) and (v_1, v_2, v_3) in PLS and sPLS (for example here when 50 genes and 30 response variables are selected) in one simulation run (results were similar for the other runs). Figure 1 shows that both PLS and sPLS highlight the "good" genes, but with no clear distinction between the group of genes with a μ_1 or a μ_2 effect for PLS in dimension 1 or 2. On the contrary, sPLS clearly selects the μ_1 effect genes on dimension 2 with heavy weights. This may be useful for the biologists who

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

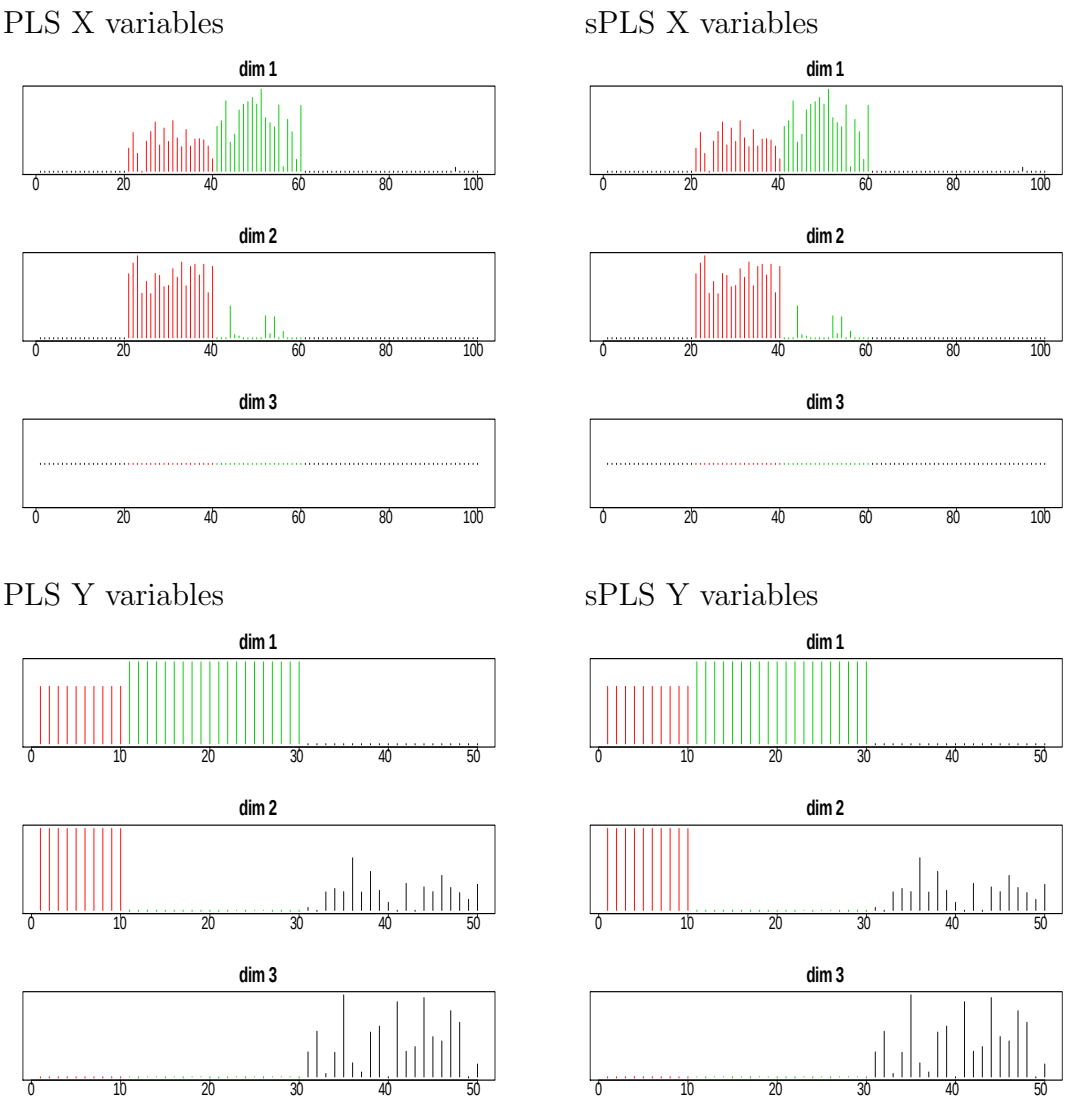


Figure 1: Absolute variable weights in the loading vectors of the PLS (left) or the sparse PLS (right, selection of 50 genes and 30 response variables) for the first 100 X variables (top) and all the Y variables (bottom) for the first three dimensions. The weights of all the X variables can be found in the supplementary material. Red (green) color stands for the variables related to the μ_1 (μ_2) effect.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Table 2: Description of the data sets.

	Liver Toxicity	Arabidopsis	Wine Yeast
# samples n	64	18	43
X	gene expressions	transcripts	transcripts
p	3116	22 810	3381
Missing values	2	0	0
Y	clinical variables	metabolites	metabolites
q	10	137	22
Missing values	0	22	0

want to clearly separate the genes related to each effect on a different dimension. For both methods, the dimension 3 did not seem to be informative. The same conclusion can be drawn for the Y variables.

If an artificial two-step selection procedure was performed in PLS, first by ordering the absolute values of the loadings and then by selecting a chosen number of variables to select 50 genes and 30 response variables for the first three dimensions, the two selections in PLS and sPLS would roughly be the same (identical for dimension 1, up to 5 different selected variables in dimension 2 and 3). This shows that sPLS simply seems to shrink the PLS loading coefficients in this simple controlled setting. However, on real data sets (see below), the difference between the two methods is genuine in terms of variable selection.

2.2 Case studies

2.2.1 Data sets

Liver toxicity study In the liver toxicity study (Heinloth et al., 2004), 4 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 the 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 (Bushel et al., 2007). There are 2 missing

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

values in the gene expression matrix.

In the original descriptive study, the authors claim that the clinical variables might not help in detecting the paracetamol toxicity in the liver, but that the gene expression information could be an alternative solution. However, in a PLS framework, it is tempting to predict the clinical parameters (Y) by the gene expression matrix (X), as performed in Gidskehaug et al. (2007).

Arabidopsis data The responses of 22810 transcript levels and 137 metabolites and enzymes (including 67 unidentified metabolites) during the diurnal cycle (6) and an extended dark treatment (6) in WT Arabidopsis, and during the diurnal cycle (6) in starch less pgm mutants, is studied (Gibon et al., 2006). The aim is to detect the changes in enzyme activities by integrating the changes in transcript levels and detecting the correlation between the different time points and the three genotypes.

According to this previous study, metabolites and enzymes are regulated by gene expression rather than vice versa. We hence assigned the transcript levels to the X matrix and the metabolites to the Y matrix. The Y data set contained 22 missing values. This data set is characterized by a very small number of samples (18).

Wine yeast data *Saccharomyces cerevisiae* is an important component of the wine fermentation process and determines various attributes of the final product. One such attribute that is important from an industrial wine-making perspective is the production of volatile aroma compounds such as higher alcohols and their corresponding esters (Nykanen and Nykanen, 1977; Dickinson et al., 2003). The pathways for the production of these compounds are not clearly delineated and much remains unknown regarding the roles and kinetics of specific enzymes. In addition, most of the key reactions in the various pathways are reversible and the enzymes involved are fairly promiscuous regarding substrate specificity (Bely et al., 1990; Ribéreau-Gayon et al., 2000). In fact, different yeast strains produce wines with highly divergent aroma profiles. The underlying genetic and regulatory mechanisms responsible for these differences are largely unknown due to the complex network structure of aroma-producing reactions. As such an unbiased, holistic systems biology approach is a powerful tool to mine and interpret gene expression data in the context of aroma compound production.

In this study, five different industrial wine yeast strains (VIN13, EC1118, BM45, 285, DV10) were used in fermentation with synthetic must, in duplicate or triplicate (biological repeats). Samples were taken for microarray

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

analysis at three key time points during fermentation, namely Day2 (exponential growth phase), Day5 (early stationary phase) and Day14 (later stationary phase). Exometabolites (aroma compounds) were also analyzed at the same time by GC-FID.

Microarray analysis was carried out using the Affymetrix platform, and all normalizations and processing was performed according to standard Affymetrix procedures. To compensate for the bias towards cell-cycle related genes in the transcriptomic data set, the data was pre-processed to remove genes that are exclusively involved in cell cycle, cell fate, protein bio synthesis and ribosome bio genesis, leaving a set of 3391 genes for a regression framework analysis, with no missing data, and $n = 43$ samples.

2.2.2 Comparisons with PLS

Comparisons with PLS will be performed on the basis of the criteria that were defined in section 1.6: Q_h^2, predictive power assessment of the model as well as insight into the variable selection and stability in the variable selections. As the main objective of this paper is to show the feasibility of the sparse approach, the three data sets will be used as illustrative examples to compare PLS and sPLS.

In this regression framework, some of the data sets are characterized by a very small number of response variables (Liver Toxicity: $q = 10$, Wine Yeast $q = 22$). In these cases, we did not deem it relevant to perform selection on the Y variables, and hence $\lambda_2^h = 0$. In the Arabidopsis data set, the selection was simultaneously performed on the X and Y data sets, as initially proposed by our approach.

Each input matrix was centered to column mean zero, and scaled to unit variance so as to avoid any dominance of one of the two data sets. Missing values were imputed with the NIPALS algorithm.

Q_h^2. We compare the Q_h^2 value with the PLS model including all variables, and the proposed sPLS model with different sparsity degrees on each dimension: selection of 50 or 150 X variables on Liver Toxicity and Wine Yeast, selection of 50 or 150 X variables coupled with the selection of 50 or 80 Y variables in Arabidopsis. The choice of the selection size was arbitrarily chosen and leave-one-out cross validation was applied for all data sets. The marginal contribution of ξ_h for each PLS/sPLS component was computed for each dimension. Figure 2 shows that the values of Q_h^2 behave differently, depending on the data set and on the PLS/sPLS approach.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

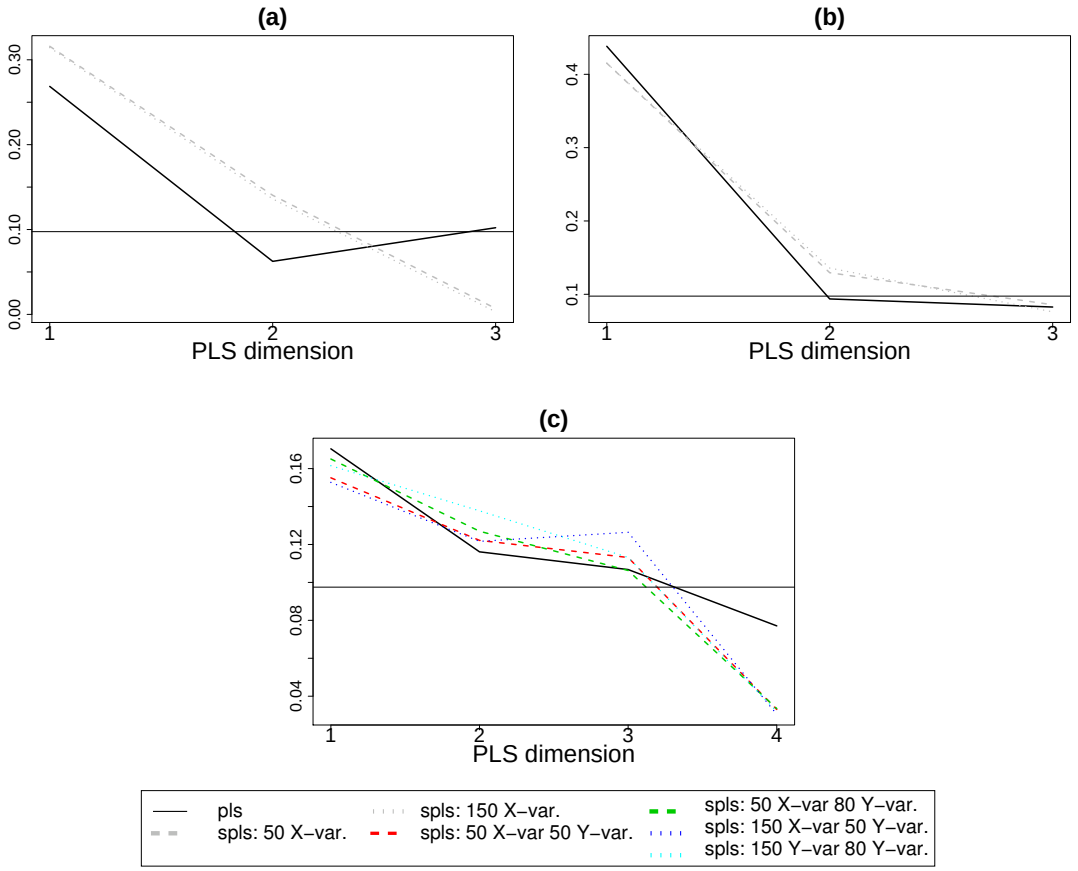


Figure 2: Marginal contribution (Q_h^2) of the latent variable ξ_h for each component in PLS and sPLS and different sparsity degrees for Liver Toxicity Study **(a)**, Wine Yeast **(b)** and Arabidopsis **(c)**. The horizontal black line indicates the threshold value of Q_h^2.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

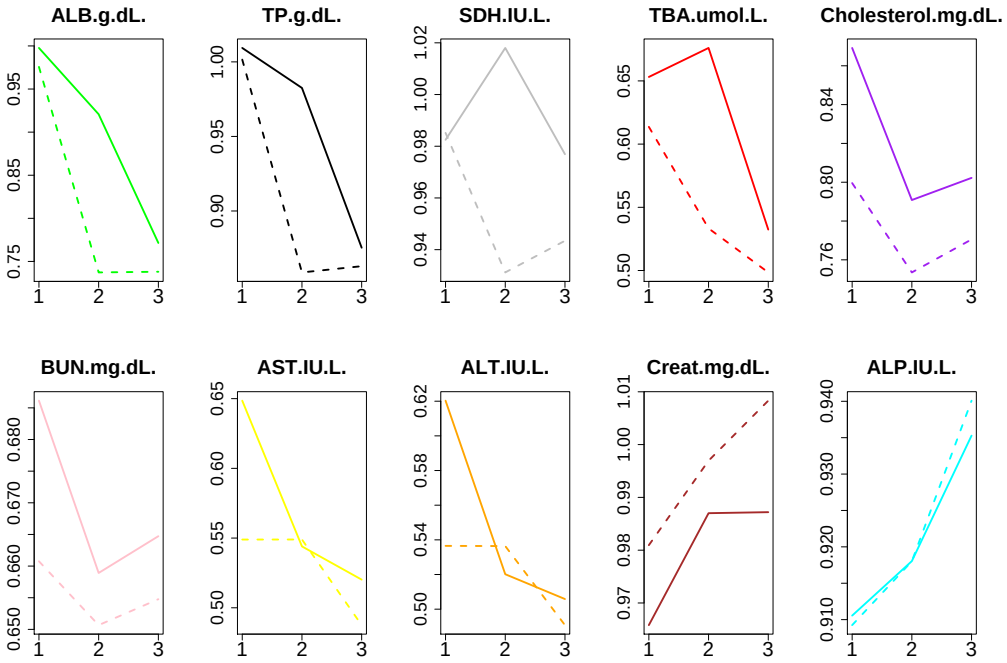


Figure 3: Liver Toxicity study: RMSEP for each clinical variable with PLS (plain line) and sPLS (dashed) for each dimension h, $h = 1 \dots 3$. Clinical variables are ranked according to their loadings in dimension 2.

In Liver Toxicity and Wine Yeast **(a) (b)**, PLS needs one less component than sPLS: 1 (2) PLS dimensions for Liver Toxicity (Wine Yeast). As already observed in section 2.1.3, sPLS would need one more dimension to fully separate the different biological effects and select the X and Y variables according to each of these effects.

In Liver Toxicity, Q^2_3 increases and becomes superior to the threshold value 0.0975. On the other hand, the Q^2_h values in any sPLS steadily decrease with h.

In Arabidopsis **(c)** which is characterized by many X variables and where a simultaneous variable selection is performed on the Y data set, the Q^2_h values differ depending on the number of variables that are selected on both data sets. However, for both methods and all sparsity degrees, the choice of $H = 3$ seems sufficient.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Table 3: Stability: ratio of the true positive variables selected in the original data sets and the bootstrap data sets over the size of each selection (100).

	Liver toxicity		Arabidopsis				Wine Yeast	
	PLS	sPLS	PLS		sPLS		PLS	sPLS
			X	Y	X	Y		
dim 1	0.735	0.739	0.332	0.895	0.377	0.893	0.596	0.598
dim 2	0.457	0.603	0.221	0.834	0.365	0.838	0.622	0.559
dim 3	0.354	0.279	0.101	0.77	0.156	0.78	0.52	0.463

Predictive ability. Figure 3 compares the RMSEP for each clinical variable in the Liver Toxicity study with PLS (no selection) and sPLS (here, selection of 150 genes). As observed in section 2.1.2, these graphics show that except for 2 clinical variables, sPLS clearly outperforms PLS. Removing some of the noisy variables in the X data set improves the prediction of most of the Y variables. In this figure, the clinical variables are ranked according to the absolute value of their weights in v_2. Hence the Y loadings have a meaning in terms of variable importance measure, as the less well explained variables *creat.mg.dL* and *ALP.IU.L* get the lowest ranks. A thorough biological interpretation would be needed here to verify if these clinical variables are relevant in the biological study.

If the clinical variables were ranked according to the next loading vector v_3, then, although the graphics would be unchanged, *creat.mg.dL* and *ALP.IU.L* would get a a higher rank (resp. rank 1 and 8). This result justify the choice of $H = 2$ for Liver Toxicity with sPLS. Similar conclusions can be drawn on the other data sets that include more Y variables.

Stability. On 10 bootstrap samples, we compare the 100 X variables and 100 Y variables (in the case of Arabidopsis) that were selected either with PLS or sPLS with respect to the same number of variables that were selected on the original (whole) data sets. Variable selection with PLS was performed in two steps: first by computing the PLS loading vectors, then by ordering the absolute values of the loading weights in decreasing order. This is similar to the simple thresholding approach proposed by Cadima and Jolliffe (1995) for PCA. The results are summarized in Table 3 and show that except for Wine Yeast in dimension 2 and 3, the sparse PLS approach seems more stable than PLS. It is not surprising to find an increased stability when the total number of variables (p and q) is rather small.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Table 4: Number of variables commonly selected in PLS (simple thresholding approach) and in sPLS when selecting 100 variables.

	Liver toxicity	Arabidopsis		Wine Yeast
	X	X	Y	X
dim 1	97	56	90	91
dim 2	56	45	82	73
dim 3	19	72	80	74

Variable selection. Table 4 highlights the actual differences between a selection performed either with PLS (in two steps) or with sPLS for the same number of variables (100 for each data set, when applicable). As expected, both selections should be similar in dimension 1, but differ greatly for the other dimensions. In particular, the selections performed in the X Arabidopsis data set differ from the very first dimension. This is due to the extremely large number of X variables ($p = 22810$), where many of the transcripts get similar weights in PLS.

2.2.3 Property of the loading vectors.

When applying sparse methods, the loading vectors may lose their property of orthogonality and uncorrelation, as it was observed with sparse PCA (Trendafilov and Jolliffe, 2006; Shen and Huang, 2008). This is not the case with sPLS. In the original PLS, no constraint is set to have $\omega_r'\omega_s = 0$, $r < s$. Hence, latent variables $(\omega_1, \ldots, \omega_H)$ from the Y data set are not orthogonal in PLS or sPLS. To remedy to this in terms of graphical representation of the samples, we propose to project $(\omega_1, \ldots, \omega_H)$ in an orthogonal basis. For the latent variables ξ, however, we always observed that $\xi_r'\xi_s = 0$ and no projection is needed for these latent variables.

3 Analysis of the wine yeast data set and biological interpretation

To begin with we will present some elements for discussion regarding the graphical representation of the latent variables (samples), which facilitate the biological interpretation. These preliminary remarks will explain some of the results that were obtained when we compared the genes selected with PLS (two-step

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Table 5: Comparison of genes selected with PLS (two-step procedure) vs. sPLS.

	PLS	sPLS
dim 1	-genes related to general central carbon metabolism -inclusion of many dubious/suspect ORFs	-GDH1: key regulator of cellular redox balance (direct influence on the main aroma producing reactions)
dim 2	-identifies "rate-limiting" enzymes in aroma metabolism	-improved coverage of transcriptional pathways
dim 3	-identifies most important alcohol and aldehydes dehyrogenase genes	-IDH1: key enzyme controlling flux distribution between aroma producing pathways and TCA cycle -NDE1: provides energy intermediates for dehydrogenase reactions

procedure) to the genes selected in the one-step procedure with sPLS. Finally we show that the sPLS selection gives meaningful insight into the biological study.

As required by the biologists who performed this experiment, 200 genes were selected on each dimension.

3.1 Biological samples

Figure 4 highlights several facts that can actually be explained by the biological experiment. The first component separated samples into time-specific clusters. This is to be expected as the particular stage of fermentation is the major source of genetic variation and the main determinant of aroma compound levels. The next most significant source of biological variation is the identity of the yeast strain. This was corroborated by the second and third components, where the samples clustered together in biological repeats of the same strain. Strains that are known to be more similar in terms of their fermentative performance also clustered closely within the time sub-groups(*i.e* EC1118 and DV10, and BM45 and 285). The VIN13 strain (which is least similar to any of the other strains in this study) showed an intermediate distribution between the latent variable axes.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

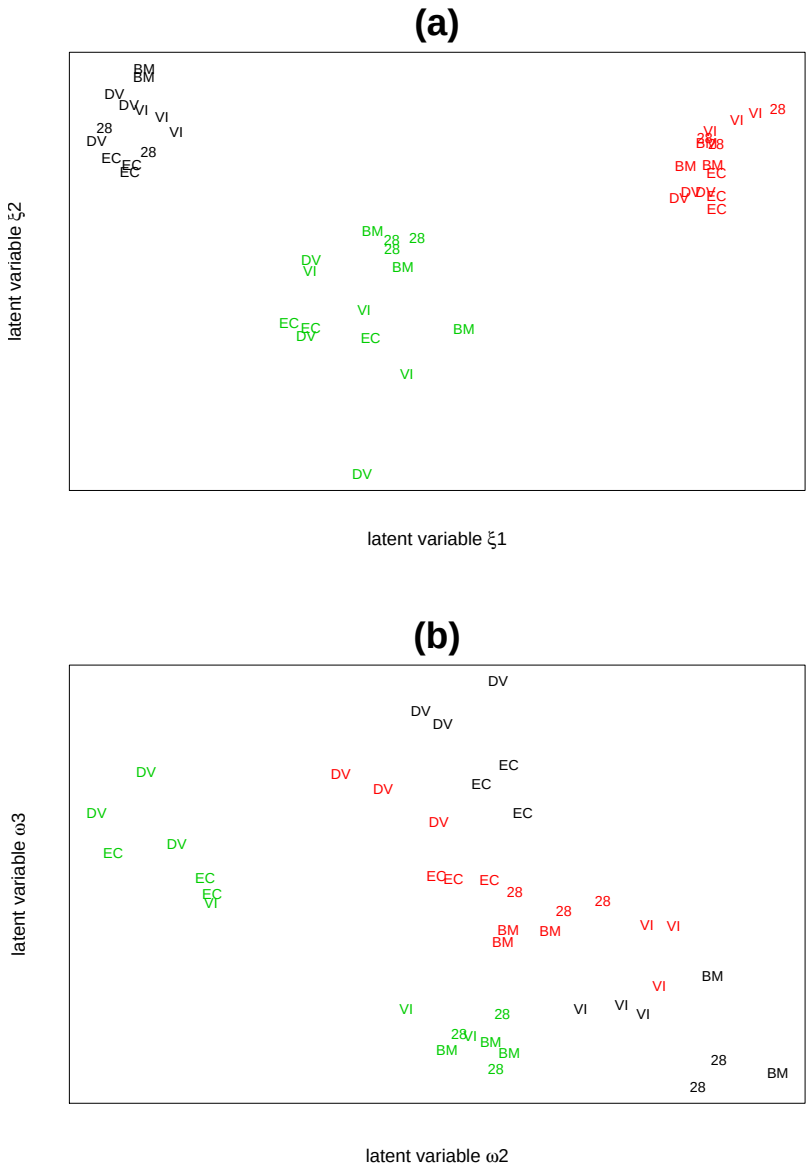


Figure 4: Wine Yeast data: graphical representation of the samples for the latent vectors (ξ_1, ξ_2) **(a)** and (ω_2, ω_3) **(b)**. Colors red, green and black stand for fermentation day 2, 5 and 14, VI = VIN13, EC = EC1118, BM = BM45, 28 = 285 and DV = DV10 .

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

3.2 Selected variables

Comparisons with PLS Table 5 presents the similarities and main differences observed between the genes which were selected either with PLS (two-step procedure) or sPLS. The striking result that we observed was the differences in the gene selections, especially in dimension 2 and 3. Overall, these dimensions were found to be more enriched for genes with proved or hypothesized roles in aroma compound production (based on pathway analysis and functional categorisation) for the sPLS rather than PLS.

Genes selected with sPLS. Figure 5 depicts the "known" or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. Indirect interactions (*i.e* missing intermediates) are indicated by dashed lines and standard reactions are indicated by solid lines. Aroma compounds (red) and other metabolic intermediates (black) are positioned at the arrow apices. Unknown enzyme activities are represented by a question mark (**?**). Gene names coding for the relevant enzymes are represented in black box format, except for those genes that were identified in the first (blue), second (purple) and third (green) components of the sPLS. This figure was constructed using GenMAPP (www.genmapp.org) and is based on KEGG pathways (www.genome.ad.jp/), in-house modifications based on available literature and MIPS (Mewes et al., 2000)/SGD (Weng et al., 2003) functional classifications.

From the figure it is clear that the sPLS outputs provided good coverage of key reactions and major branches of the aroma production pathways (for the areas of metabolism with known reactions and enzymes). The first component identified mostly genes that are involved in reactions that produce the key substrates for starting points of the pathways of amino acid degradation and higher alcohol production. Amino acid metabolism is also a growth stage-specific factor (linked to fermentative stage), which is supported by the observations discussed in section 3.1. Most of the crucial "rate limiting" enzymes (PDC2, ALD2, ALD3, LEU1) were identified by the second component. In total, the highest number of relevant genes were identified by the third component. Genes in this component were also interesting from the perspective that they only have putative (but unconfirmed) roles to play in the various pathways where they are indicated in the figure. Associations between genes with putative functional designations (based on homology or active site configuration) and aroma compounds in the lesser annotated branches of aroma compound production provide opportunities for directed research and the formulation of novel hypothesis in these areas.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

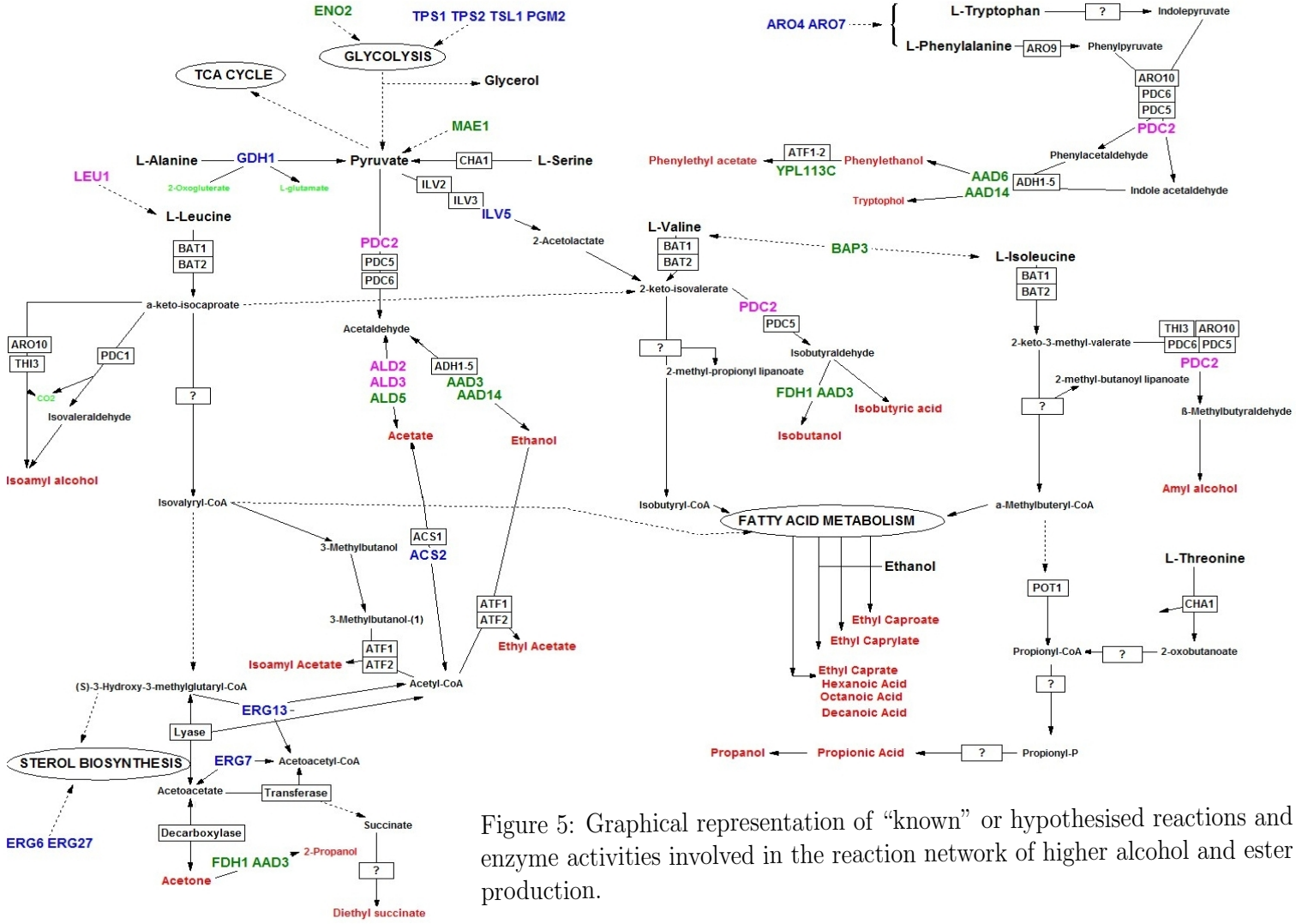


Figure 5: Graphical representation of "known" or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/1/15 3:33 AM

Further analysis to be done. An attractive way of representing variables is to compute the correlation between the original data set (X and Y) and the latent variables ($\xi_1, \ldots \xi_H$) and ($\omega_1, \ldots \omega_H$), as it is done with PCA or CCA. The selected variables are then projected on a correlation circle. This enable the identification of known and unknown relationships between the X variables, the Y variables, and more importantly between both types of omics data. Of course these relationships will then need to be biologically assessed with further experiments, and will constitute a next step of our proposed analysis.

4 Conclusion

We have introduced a general computational methodology that modifies PLS, a well known approach that has proved to be extremely efficient in many data sets where $n << p + q$. In the sparse version, variable selection is included with Lasso penalization in order to be more useful and applicable to biological problems. Validation of the sparse PLS approach has been performed not only on a simulated data set but also on real data sets and compared with PLS. The simulation study showed that sPLS selected the relevant variables which were governed by the known latent effects. Application to real data sets showed that this built-in variable selection procedure improved the predictive ability of the model, differed from PLS from dimension 2 onwards and seemed more stable. Compared to PLS, sPLS seemed to separate each latent biological effect on a different dimension and accordingly selected the variables governed by each effect. This result will help biologists to identify relevant variables linked to each biological condition.

Our proposed algorithm is fast to compute. Like any sparse multivariate method, sPLS requires the addition of penalization parameters. The tuning of these two parameters can simply be performed either by estimating the prediction error with cross validation when the number of samples permits it, or by choosing the variable selection size when n is too small-a useful option for the biologist. The gain of penalizing, and hence by selecting variables is validated in a typical biological study aimed at integrating gene expression and metabolites in wine yeast. We provide a thorough biological interpretation and show that the sPLS results are extremely meaningful for the biologist, compared to a PLS selection. This preliminary work undoubtedly brings more insights into the biological study and will suggest further biological experiments to be performed.

Our sparse PLS was mainly studied from an empirical point of view and

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

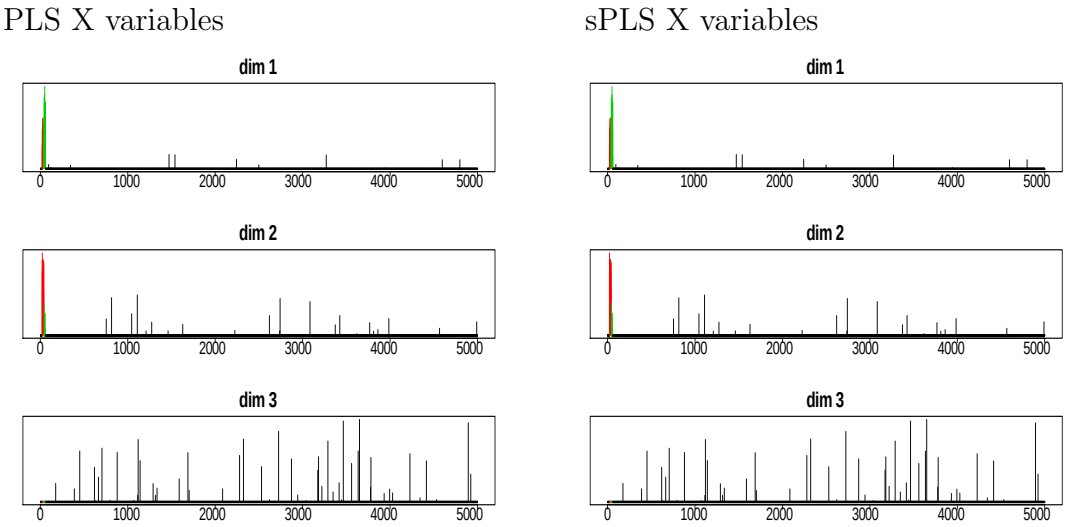


Figure 6: Supplemental figure: absolute variable weights in the loading vectors of PLS (left) or sparse PLS (right) for the 5000 X variables for the first three dimensions. Red (green) color stands for the variables related to the μ_1 (μ_2) effect.

showed very promising results. Further theoretical justifications are needed for a deeper understanding of the performance of this approach. As integrating omics data is an issue that may soon be widely encountered in most high throughput biological studies, we believe that our sparse PLS will provide an extremely useful tool for the biologist in need of analyzing two-block data sets. It indeed provides an easier interpretation of the resulting variable selections.

Remark 1. sPLS can be applied with other thresholding rules, such as the hard thresholding function. When the latter rule was used (not shown) the sparse PLS was similar to the PLS applied with a simple thresholding approach (two-step procedure) and hence did not yield relevant results. Note also that other penalty functions could be considered, such as Elastic Net (Zou and Hastie, 2005) or the Bridge penalties (Frank and Friedman, 1993) to extend this sparse PLS approach.

Remark 2. Another variant in our sparse PLS approach can be considered in step (g) of the proposed algorithm in section 1.4 by deflating the Y matrix in a symmetric manner: $Y_h = Y_{h-1} - \omega_h e'_h$. In this case, we are in a canonical

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

framework and the aim is to model a *reciprocal relationship* between the two sets of variables. The lack of statistical criteria in this setting (as we are not in a predictive context) would require a thorough biological validation of the approach, rather than a statistical validation, and will constitute the next step of our research work.

Availability The code sources of sparse PLS (in `R`, the Comprehensive R Archive Network, http://cran.r-project.org/) can be available upon request to the corresponding author. An `R` package is currently being implemented.

References

Bely, M., Sablayrolles, J., and Barre, P. (1990). Description of Alcoholic Fermentation Kinetics: Its Variability and Significance. *American Journal of Enology and Viticulture*, 41(4):319–324.

Boulesteix, A. (2004). PLS Dimension Reduction for Classification with Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1075.

Boulesteix, A. and Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling*, 2(23).

Burnham, A., Viveros, R., and Macgregor, J. (1996). Frameworks for latent variable multivariate regression. *Journal of chemometrics*, 10(1):31–45.

Bushel, P., Wolfinger, R. D., and Gibson, G. (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology*, 1(15).

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52:1181–1191.

Cadima, J. and Jolliffe, I. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214.

Chun, H. and Keles, S. (2007). Sparse Partial Least Squares Regression with an Application to Genome Scale Transcription Factor Analysis. Technical report, Department of Statistics, University of Wisconsin, Madison, USA.

Culhane, A., Perriere, G., and Higgins, D. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Datta, S. (2001). Exploring relationships in gene expressions: A partial least squares approach. *Gene Expression*, 9(6):249–255.

de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263.

De Jong, S. and Ter Braak, C. (1994). Comments on the PLS kernel algorithm. *Journal of chemometrics*, 8(2):169–174.

Dickinson, J., Salgado, L., and Hewlins, M. (2003). The Catabolism of Amino Acids to Long Chain and Complex Alcohols in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 278(10):8028–8034.

Dolédec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, 31(3):277–294.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455.

Dray, S., Pettorelli, N., and Chessel, D. (2003). Multivariate Analysis of Incomplete Mapped Data. *Transactions in GIS*, 7(3):411–422.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

Gibon, Y., Usadel, B., Blaesing, O., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. *Genome Biology*, 7:R76.

Gidskehaug, L., Anderssen, E., Flatberg, A., and Alsberg, B. (2007). A framework for significance analysis of gene expression data using dimension reduction methods. *BMC Bioinformatics*, 8(1):346.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389–422.

Heinloth, A., Irwin, R., Boorman, G., Nettesheim, P., Fannin, R., Sieber, S., Snell, M., Tucker, C., Li, L., Travlos, G., et al. (2004). Gene Expression Profiling of Rat Livers Reveals Indicators of Potential Adverse Effects. *Toxicological Sciences*, 80(1):193–202.

Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2(3):211–228.

Krämer, N. (2007). An overview of the shrinkage properties of partial least squares regression. *Computational Statistics*, 22(2):249–273.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Lê Cao, K.-A., Gonçalves, O., Besse, P., and Gadat, S. (2007). Selection of biologically relevant genes with a wrapper stochastic algorithm. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 1.

Lorber, A., Wangen, L., and Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics*, 1(19-31):13.

Mevik, B.-H. and Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2).

Mewes, H., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuler, C., Strocker, S., and Weil, B. (2000). Mips: a database for genomes and protein sequences. *Nucleic Acids Research*, 28:37–40.

Nykanen, L. and Nykanen, I. (1977). Production of esters by different yeast strains in sugar fermentations. *J. Inst. Brew*, 83:30–31.

Pihur, V., Datta, S., and Datta, S. (2008). Reconstruction of genetic association networks from microarray data: A partial least squares approach. *Bioinformatics*.

Ribéreau-Gayon, P., Dubourdieu, D., Donèche, B., and Lonvaud, A. (2000). *Handbook of Enology*, volume 1. John Wiley and Sons.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99:1015–1034.

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions Technip.

Thioulouse, J., Chessel, D., Dolédec, S., and Olivier, J. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing*, 7(1):75–83.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Trendafilov, N. and Jolliffe, I. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics and Data Analysis*, 50(1):242–253.

Trygg, J. and Wold, S. (2003). O2-pls, a two- block (x-y) latent variable regression (lvr) method with an integral osc filter. *Journal of Chemometrics*, 17:53–64.

Umetri, A. (1996). SIMCA-P for windows, Graphical Software for Multivariate Process Modeling. *Umea, Sweden*.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM

Waaijenborg, S., de Witt Hamer, V., Philip, C., and Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1):3.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.

Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S., Engel, S., Fisk, D., Hong, E., et al. (2003). Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Research*, 31(1):216.

Wold, H. (1966). *Multivariate Analysis*. Academic Press, New York, Wiley, krishnaiah, p.r. (ed.) edition.

Wold, S., Eriksson, L., Trygg, J., and Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Technical report, Umea University.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320.

Brought to you by | University of Queensland - UQ Library
Authenticated
Download Date | 12/11/15 3:33 AM