Sparse PLS: Variable Selection when Integrating Omics data

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

Sparse PLS: Variable Selection when Integrating Omics data

Kim-Anh Lê Cao[1,2*], Debra Rossouw[3], Christèle Robert-Granié[2], Philippe Besse[1]

Abstract

Recent biotechnology advances allow the collection of multiple types of omics data sets, such as transcriptomic, proteomic or metabolomic data to be integrated. The problem of feature selection has been adressed several times in the context of classification, but has to be handled in a specific manner when integrating data. In this study, we focus on the integration of two types of data sets that are measured on the same samples. Our goal is to combine integration and variable selection in a one-step procedure using the good properties of PLS to facilitate the biologists interpretation. A novel computational methodology called sparse PLS is introduced, with two variants depending on the modelling or predictive purpose of the analysis, to deal with these newly arisen problems. The sparsity of our approach is obtained by Lasso penalization of the loading vectors in a SVD PLS version.

Sparse PLS is shown to be effective and biologically meaningful. Comparisons with classical PLS are performed on real data sets and a thorough biological interpretation of the results obtained on one data set is provided. We show that sparse PLS not only benefits from the attractive stability property of PLS but also provide a valuable variable selection tool for high dimensional data sets.

1 Introduction

1.1 Motivation

Recent advances in technology enable the monitoring of an unlimited quantity of data from various sources. These data are gathered from different analytical platforms and allow their integration among different types, such as transcriptomic, proteomic or metabolomic data. This integrative biology approach enable to understand better some underlying biological mechanisms and interaction between functional levels if one succeeds in incorporating the several omics types of data, characterized by many variables but not necessarily many samples or observations. In this highly dimensional setting, the selection of genes, proteins or

[*]to whom correspondence should be addressed: Kim-Anh.Le-Cao@toulouse.inra.fr

[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France

[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France

[3]Institute for Wine Biotechnology, University of Stellenbosch, Stellenbosch, South Africa

metabolites is absolutely crucial to overcome computational limits (from a mathematical and statistical point of view) and to facilitate the biological interpretation. Hence our quest of sparsity is motivated by the biologists needs, who want to separate the useful information related to the study from the non useful information, due to experiment inaccuracies. The resulting variable selection might also enable a feasible biological validation with a reduced experimental cost.

In this paper, we especially focus in the integration context, which is the main goal of omics data. For example, one biological study might aim at explaining the q metabolites by the p transcripts, that are measured on the same n samples. In this typical case, $n << p+q$.

Here we propose a sparse version of the PLS, that aims at combining selection *and* modelling in a one-step procedure for such problems. Our sparse PLS is based on Lasso regression (Tibshirani, 1996) and is obtained by penalizing a sparse SVD (Shen and Huang, 2007), using a specific PLS-SVD variant (Lorber et al., 1987).

The sparse PLS that we propose deals with integration problems, that cannot be solved with usual feature selection approaches proposed in classification or discrimination studies where there is only one data set to analyse. Hence, multiple testing that looks for differentially expressed genes does not apply here, as well as other classification methods that were applied to transcriptomic data sets. In this latter case, many authors (among them: Guyon et al. 2002; Lê Cao et al. 2007) have applied feature selection methods to microarray data and have been proved to bring biologically meaningful genes lists. However, in our context, the feature selection aim has to be integrated with modelling, and very few approaches have been proposed to deal with these newly arisen problems, especially in a one-step procedure.

In this study, we focus on the integration of two types of data matrices X ($n \times p$) and Y($n \times q$) that are measured on the same observations. In this two-block data sets setup, our aim is to *model* the relationships between the two types of variables, or to *predict* one group of variables from the other group. Several approaches that seek linear combinations of both groups of variables can answer this biological problem. However, they are often limited by collinearity or ill posed problems, that require regularization techniques, such as l_1 (Lasso) or l_2 (Ridge) penalizations.

Among them, Canonical Correlation Analysis (CCA, Hotelling 1936) aims at maximizing the correlation between linear combinations of the X variables and Y variables. However, CCA becomes very unstable when the number of variables becomes large, and when $n << p+q$, the exact computations of the inverses $(X'X)^{-1}$ and $(Y'Y)^{-1}$ is not feasible. Vinod (1976) proposed to regularize CCA (rCCA) with an l_2 penalization so as to overcome this limit and González et al. (2008) proposed to tune the penalization parameters with a cross validation strategy. Their application to biological data showed relevant results. However, when the number of variables becomes too large, the graphical outputs are difficult to interpret, as no selection procedure is performed. Our experience showed that adding a lasso parameter to CCA in fact leads to very unstable results. Waaijenborg et al. (2008) recently circumvent the issue by adapting the Elastic Net (Zou and Hastie, 2005) that combines l_1 and l_2 penalizations to CCA in a regression framework, leading to sparse canonical factors. This penalized CCA does not optimize the canonical correlation criteria of the original CCA, but maximize the penalized version of the canonical correlation instead. This results in a non monotonic decreasing trend in canonical correlation, and hence a difficult choice to make concerning the number of canonical variate pairs. On a Glioma data set, the authors show that some of

the gene selections and gene copy number selections coming from the first 20 canonical pairs might bring relevant results, and the first pair highlighted one gene known as being implied in the biological study.

Partial Least Squares regression (PLS, Wold 1966) is a well known regression technique, mostly applied in chemometrics. Its stability property faced to collinear matrices gives PLS a clear superiority to CCA, multiple linear regression, ridge regression or other regression techniques. Furthermore, since Wold 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 a solution for almost any problem. We will describe and discuss some of these variants in this study.

PLS has 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, Doledec and Chessel 1994) from which PLS is a particular case, in a cross platform comparison in microrray data.

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

In this paper, we propose a sparse PLS approach that combines both integration and variable selection, in a one-step strategy. We also provide two variants, in case of the modelling (exploratory) or the predictive aim of the analysis. Several ways of tuning of the two penalization parameters and choosing the PLS dimension are proposed, such as cross validation in the regression case. But these techniques are often limited when confronted to real data sets characterized by n very small. This is why we prefered to show instead that sparse PLS is applicable on real data sets. With the help of biologists, we show how useful and meaningful sparse PLS is, instead of proving it on non realistically simulated data sets. Indeed, when developping tools for biological data that require applicability, we must be aware that often, statistical criteria cannot stand in this extreme setting, and that we should mostly rely on biological interpretation, for example to guide the number of PLS components.

1.2 Outline of the paper

A brief introduction to PLS will be given, before describing the sparse PLS method and its two variants. We detail how to add sparsity to PLS with a Lasso penalization (Tibshirani, 1996) combined to SVD computation (Shen and Huang, 2007). We then assess the validity of the approach on four real data sets, compare and discuss the results with a classical PLS approach. We also provide a full biological interpretation of the results obtained on a typical integrative study of wine yeast, that combines transcripts and metabolites. We show how sparse PLS highlights the most essential transcripts that are meaningfully related to the metabolites.

2 Methods

2.1 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: $(\xi_1, \xi_2 \dots \xi_H)$, $(\omega_1, \omega_2 \dots \omega_H)$, that are n-dimensional vectors, and associated loadings: $(u_1, u_2 \dots u_H)$, $(v_1, v_2 \dots v_H)$, that are respectively p and q- dimensional vectors, to solve the following optimization problem (Burnham et al., 1996):

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

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

Many PLS variants exist depending on the way X and Y are deflated, either in a symetric ("PLS-mode A") or asymetric way ("PLS2") (Tenenhaus, 1998; Wegelin, 2000), and the models will consequently differ.

For example, in case of *regression mode* (asymetric), the models of X- and Y-space 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, 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 $d_h = Y_{h-1}^T \xi_h / (\xi_h' \xi_h)$, and ε_1 ($n \times p$) and ε_2 ($n \times q$) are the residual matrices, $h = 1 \dots H$.

In case of *canonical mode* (symetric), X and Y are in contrary deflated in a symetric way and the models become:

$$X = \Xi C^T + \varepsilon_1 \qquad Y = \Omega E^T + \varepsilon_2$$

where Ω ($n \times H$) is the matrix of PLS components ω_h and the column vectors of E are defined as $e_h = Y_{h-1}^T \omega_h / (\omega_h' \omega_h)$, $h = 1 \dots H$.

Another PLS alternatives exist depending if X and Y are deflated separately or directly using the crossproduct $M = X^T Y$ and the SVD decomposition. We will discuss these various approaches in sections 2.2 and 2.4.

Note that in any case, all PLS variants are equivalent during the computation of the first dimension.

2.2 SVD decomposition and PLS-SVD

We recall the SVD decomposition and the principle of the PLS-SVD approach, that will be useful for understanding our sparse PLS approach.

2.2.1 Singular Value Decomposition

Any real r-rank matrix M $(p \times q)$ can 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 $(k = 1 \dots r)$ are called the singular values. The singular values are equal to the square root of the eigenvalues of the matrices $M^T M$ and MM^T. One interesting property that will be used in our sparse PLS method is that the columns vectors of U and V, noted $(u_1, \dots, u_r)$ and $(v_1, \dots, v_r)$ (resp. called left and right singular vectors) correspond to the PLS loadings of X and Y if $M = X^T Y$.

2.2.2 PLS-SVD

In PLS-SVD, the SVD decomposition of $M = X^T Y$ is performed only once, and 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 however lead to non mutually orthogonal latent variables, a property of PLS2 $(\xi_s' \xi_r = 0, r < s)$ and PLS-mode A $(\xi_s' \xi_r = 0 \; and \; \omega_s' \omega_r = 0, r < s)$.

2.3 Lasso penalization

Shen and Huang (2007) proposed a sparse PCA approach using the SVD decomposition of $X = U\Delta V^T$ by penalizing the PCA loading vector v_k. 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. Among different penalty functions proposed, we can consider the L_1, also called Lasso penalty (Tibshirani, 1996).

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 $g(y) = sign(y)(|y| - \lambda)_+$ is the Lasso penalty function.

In our particular PLS case, we are interested in penalizing both loadings vectors u_k and v_k to perform variable selection in both data sets. Indeed, one interesting property of PLS is the direct interpretability of the loadings vectors as a measure of the relative importance of the variables in the model (Wold et al., 2004). Our optimization problem becomes:

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

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

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

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

The sparse PLS algorithm is detailed in next section.

2.4 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 compute separately X_h and Y_h, then to decompose at each step $\tilde{M}_h = X_h^T Y_h$ 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. Furthermore, this will allow us to consider two variants of our sparse PLS, depending on the way the matrix Y is deflated.

If Y is deflated in an asymetric way, we are in the regression framework where the X variables explain the Y variables. Whereas if Y is deflated in a symetric way (canonical mode), the aim is to model the relationships between the X and the Y variables, and hence highlight interactions between the two sets of variables.

In our approach, the SVD decomposition will provide a useful tool for selecting variables from each of the two-blocks data. We now detail the sparse PLS algorithm (*sPLS*) with its two deflation variants, based on the iterative PLS algorithm (Tenenhaus, 1998).

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. $u_{new} = g_{\lambda_2}(\tilde{M}_{h-1} v_{old})$, normalize u_{new}
      ii. $v_{new} = g_{\lambda_1}(\tilde{M}_{h-1}^T u_{old})$, normalize 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) $c_h = X_{h-1}^T \xi_h / \xi_h' \xi_h$
$d_h = Y_{h-1}^T \xi_h / \xi_h' \omega_h$
$e_h = Y_{h-1}^T \omega_h / \omega_h' \omega_h$

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

(g) $\begin{cases} \text{Regression mode:} & Y_h = Y_{h-1} - \xi_h d_h' \text{(asymetric)} \\ \text{Canonical mode:} & Y_h = Y_{h-1} - \omega_h e_h' \text{(symetric)} \end{cases}$

Note that in the case where there is no sparsity constraint ($\lambda_1 = \lambda_2 = 0$) we obtain the same results as in a classical PLS (PLS-mode A or PLS2).

2.5 Missing data

When dealing with biological data, it is very common to be confronted to missing data. In order not too loose too much information, an interesting approach to substitute each missing data with a value can be 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 block data set. The principle of the method relies on iteratively computing PLS loadings $(\xi_1, \ldots, \xi_H)$ and components $(c_1, \ldots, c_H)$ as slopes of least squares lines passing through the origin on the available data. Missing data x_{ij} are then estimated with these simple regressions: $\hat{x}_{ij} = \sum_{l=1}^{h} \xi_{li} c_{lj}$. 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%), see for example Dray et al. (2003). NIPALS is now implemented in the `ade4` package.

2.6 Tuning criteria and evaluation

There are two sorts of parameters to tune in sPLS: the lasso penalizations λ_1 and λ_2 for each PLS dimension and the choice of the dimension H. Here we propose different strategies to deal with this issue.

2.6.1 Lasso penalization

Explained variance. In any sparse approach that has been developped for high throughput data (essentially sparse PCA variants, see Jolliffe et al. 2003; Zou and Hastie 2005; Shen and Huang 2007), it is worth noting that the explained variance on each component is expected to be lower than in the non sparse version. This is mainly due to the fact that some noise or technical effects are actually part of the variance. Hence a tuning criteria based on how much the explained variance decreases compared to the non sparse version has often been proposed (Zou and Hastie, 2005; Shen and Huang, 2007). For example one could choose the penalization parameters so that the explained variance does not drop more than 10 % on the chosen H component compared to the non penalized version of PLS (Shen and Huang, 2007).

Degree of sparsity. The choice of the penalty parameters can also be empirically guided through the sparsity chosen by the user. Indeed in biological data sets, many omics data are still unknown (*e.g* associated function, annotation) and a too small selection of those might not allow the biologist to assess the results. This is why he may prefer instead to choose the

number of non zero components in each loading vector u_h, v_h or in both, for each dimension h.

2.6.2 Choice of PLS dimension

Marginal contribution of the latent variable ξ_h. In the case of regression mode, 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. Here, as the number of samples n is usually small, we propose to use leave-one-out cross validation (loo-cv). Q_h^2 is computed on all the 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 i sample, we can predict $\hat{y}_{hi}^k = x_{hi}\hat{B}_{h(-i)}^k$.

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

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

Of course in the case of the canonical mode, this latter rule cannot exist. Then as in any other multivariate method (CCA, PCA), the choice of H has to be made empirically by looking at the singular values or by setting a threshold to the cumulative percentage of explained variance.

In any case, the choice of the PLS dimension remains an open question that has been mentionned by several authors (Mevik and Wehrens, 2007; Boulesteix, 2004). In our particular biological context, and to facilitate graphical representations, which is one of the objectives of our sparse approach, we advise to take $H \leq 3$. Indeed our results (see below) show that very relevant information can already be extracted from 3 dimensions.

2.6.3 Evaluation

RMSEP In the case of regression mode, 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 to use the Q_h^2 criterion for this issue, we propose instead to use the RMSEP criterion as a way of evaluating the predictive power for each Y variable between the original non-penalized PLS and the sPLS, with several penalization parameters.

Note that the Q_h^2 criteria is closely related to RMSEP ($= PRESS_{kh}$), but it gives a more general insight of the PLS, whereas the RMSEP requires to be computed for each variable k in Y.

Once again this criterion cannot be computed in case of canonical mode and only a biological interpretation can allow to assess the results with this deflation mode.

2.7 The choice of the deflation mode

In the following sections, we propose to apply one of the two proposed deflation modes and show that only one of each variant can be adequate to the biological study.

The regression mode can be applied when there is an a priori knowledge about which group of variables implies the other group. This is often the case with omics data, as the three main functional levels (transcriptomic, proteomic and metabolomic) constitute known successive events in the cell life. Hence, considering that trancripts or genes $\leftarrow$ metabolites $\leftarrow$ proteins, a regression mode to model the predictive ability of *e.g* metabolites given transcripts is perfectly adapted.

On the contrary, the canonical mode can be applied when there is no a priori knowledge about the relationship between the two groups of variables, hence adopting an exploratory approach, or if we know that the two groups of variables are reciproqually related. We will provide an illustrative example where this is the case.

3 Validation studies

The evaluation of any statistical approach is usually performed with simulated data sets. However, in the context of biological data, simulation is a difficult exercise as one has to take into account technical effects that are not even easily identifiable on the real data sets. Moreover, a mathematical underlying model is needed to simulate the data, and the closer the model to the evaluated approach, the better the results. Many authors proposed algorithms to "realistically" simulate a microarray data set and demonstrate the complexity of the task (see for example Hein et al. 2005; Nykter et al. 2006). Because of this issue, it is becoming more and more frequent to directly evaluate the proposed approach on real data sets and either compare the results with other closely related approaches or show that the results are biologically relevant (Culhane et al., 2003; Johansson et al., 2003; Boulesteix, 2004; Steinfath et al., 2008).

In our case, where we want to maximize the covariance between two high dimensional data sets, the simulation procedure is even more difficult, as Nguyen and Rocke (2004) underlined. Hence, we decided to assess the results of our sparse approach with real-world data sets, as our main aims are to show that our approach is applicable on biological data sets and that it may give potentially relevant results. To show what we gain by applying our sparse approach, we will compare sparse PLS to the original PLS from a statistical but also from a biological point of view.

We hence analysed four real data sets with various complexities in terms of biological goals and number of variables in each set (see table 1). We describe the various aims that led to their study and explain why a PLS approach with an appropriate deflation mode would be adapted in these cases. We will then compare our sparse PLS to the original PLS and finally we will provide a detailed biological interpretation for one of the data set, to show that we answer the biological question.

Table 1: Description of the data sets.

	Liver Toxicity	Arabidopsis	NCI60	Wine Yeast
# samples n	64	18	60	43
X	gene expr	transcript	cDNA transcript	transcript
p	3116	22 810	5643	3381
Missing values	2	0	0	0
Y	clinic	metabolite	Affymetrix probes	metabolite
q	10	137	1517	22
Missing values	0	22	0	0
Deflation mode	regression	regression	canonical	regression

3.1 Description of real data sets

3.1.1 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) dose 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 variables containing markers for liver injury are available for each object and numerically measure the serum enzymes level. The expression data are arranged in a matrix X of $n = 64$ objects and $p = 3116$ expression levels after normalization and preprocessing (Bushel et al., 2007). There are 2 missing values in the gene expression matrix.

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

3.1.2 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 starchless pgm mutants, is studied (Gibon et al., 2006). Note that an actual selection of 82 transcript known as coding for the known metabolites and enzymes was performed in the referenced paper. We prefered instead to rely on the original data set, which framework is more adapted to our approach. The aim is to detect the change of enzyme activities by integrating the changes in transcript levels and detect the correlation between the different time points and the 3 genotypes.

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

3.1.3 NCI60

In this study, 60 cancer cell-line are considered, including cell-lines of cancers of colorectal (7), renal (8), ovarian (6), breast (8), prostate (2), central nervous system origin (9), leukemias (6) and melanomas (8) and different gene expression are measured on two different array support : cDNA (Scherf et al., 2000; Ross et al., 2000) and Affymetrix (Staunton et al., 2001). Pre-processing was performed by Culhane et al. (2003) and missing data were imputed with k-nearest neighbours by these authors. This cross-platform comparison is the perfect framework for canonical deflation mode, although we are not able yet to provide a thorough biological interpretation of the results. The cDNA data set ($= X$) contains 5643 genes and the Affymetrix data set ($= Y$) 1517 probe sets.

3.1.4 Wine Yeast data set

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 fermentations with synthetic must, in duplicate or triplicate (biological repeats). Samples were taken for microarray 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 analysed at the same time by GC-FID.

Microarray analysis was carried out using the Affymetrix platform, and all normalisations 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 biosynthesis and ribosome biogenesis, leaving a set of 3391 genes for a regression framework analysis, with no missing data, and $n = 43$ samples.

3.2 Comparison of the PLS variants

As our sparse PLS can be considered as a PLS variant, but with a variable selection procedure included, we propose to focus only on the PLS approach and compare our two sparse variants with an appropriate deflation mode of Y. This will be done in order to assess the validity of our approach.

Comparison will be performed in terms of criteria defined in section 2.6: variance explained

by PLS components for each variant, predictive power assessment of the model (regression mode only) and also graphical outputs that are the main tool to facilitate the biological interpretation. As the main objective of this paper is to show the feasability of the sparse approach, the four data sets will be used as illustrative examples to compare PLS and sPLS, and we will provide the biological interpretation of only one data set (Wine Yeast). In the case of NCI60 where the canonical mode was applied, the validation of the results that we propose would require much input from the biological side that will not be presented in this study.

Note that most of our biological data sets fitting in the regression framework are characterized by q very small (Liver Toxicity: $q = 10$, Wine Yeast $q = 22$). In these cases, we did not judge relevant to perform selection on these Y variables, and hence $\lambda_2^h = 0$. In the other data set Arabidopsis, the selection was performed on the Y data set.

4 Results and discussion

Each input matrix has been centered to column mean zero, and scaled to unit variance so as to avoid any dominance of one of the two-block data sets. Missing values have been imputed with the NIPALS algorithm. In order to compare different sparsity degrees in sPLS, a selection of 50, 100, 150 and 200 X variables on each dimension is proposed and the different validation criteria are computed.

4.1 Variance explained

Figure 1 compares the cumulative proportion of explained variance for $H = 3$ components (here H is arbitrarily chosen for illustrative purposes) for PLS and various sPLS in Arabidopsis (regression mode, top) and in NCI60 (canonical mode, bottom, for the cDNA (left) and Affymetrix (right) data sets). The same trend as in Arabidopsis was observed in Liver Toxicity and Wine Yeast in regression mode.

There seems to be a decrease of approximatively 1-5% of the proportion of cumulative explained variance from PLS to sPLS in any data set, which is not much. The fact that the explained variance seems to be similar for any sparsity degree in sPLS, suggests that this actual loss of information in the data set may be due to noise only rather than meaningful information. The low percentage in NCI60 shows the complexity of the data set.

4.2 Q_h^2

We computed the marginal contribution of ξ_h for each PLS/sPLS component, for H arbitrarily set to 3. Figure 2 shows that the values of Q_h^2 behave differently, depending on the data set and on the PLS/sPLS approach.

In Liver Toxicity and Yeast **(a) (b)**, PLS seems to need one less component than sPLS (H = 1 for Liver Toxicity and 2 for Yeast with PLS). This conclusion should however be taken with caution in Liver Toxicity, as Q_3^2 increases and becomes superior to the threshold value 0.0975. On the other hand, the Q_h^2 values in any sPLS steadily decreases with h.

In Arabidopsis **(c)** that is characterized by many X variables, the number of PLS components will need to be much greater than in the sPLS, for which $H = 3$ seems sufficient. This graphic probably shows the actual limit of PLS when p or q becomes extremely large. Note

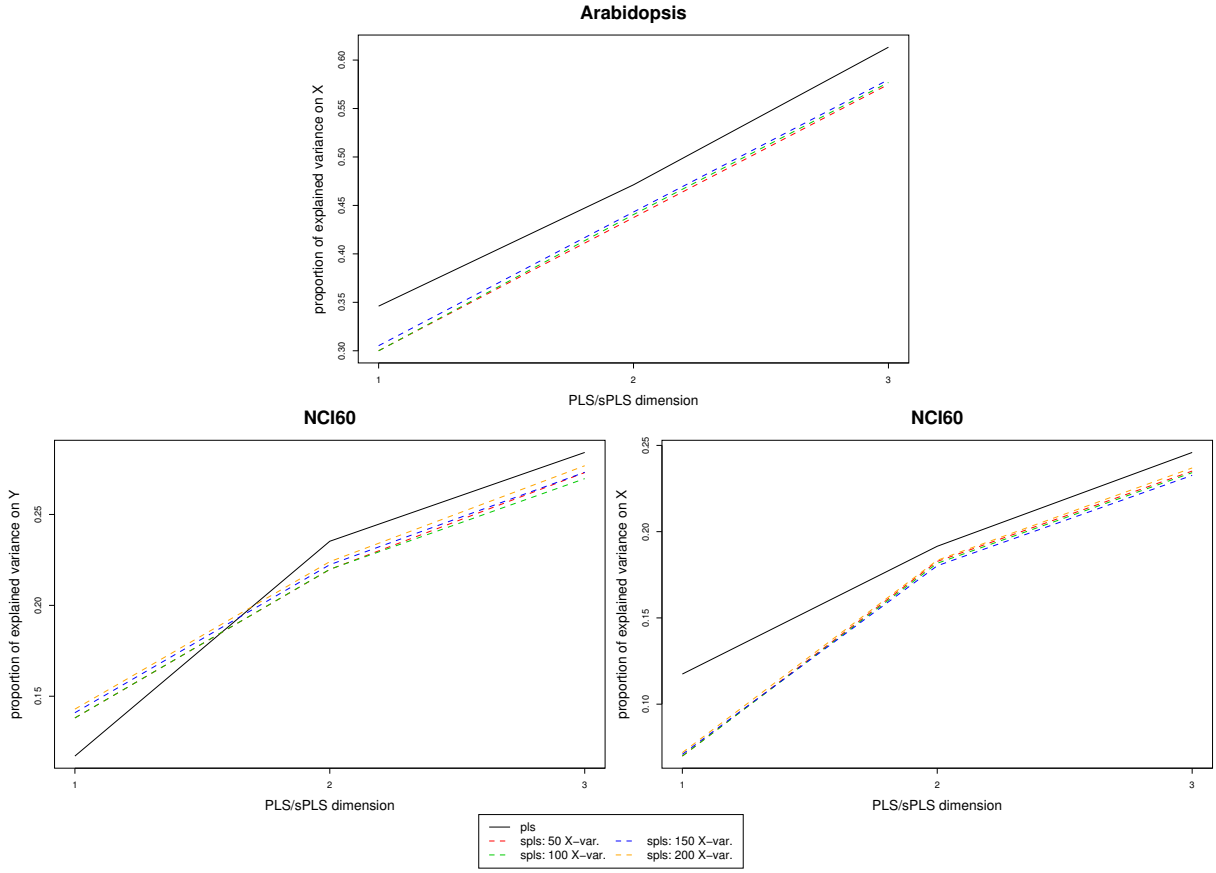

Figure 1: Cumulative proportion of explained variance for 3 components in PLS and sPLS for different sparsity degrees in Arabidopsis (regression mode, top) and NCI60 (canonical mode, bottom, for the cDNA (left) and Affymetrix (right) data sets).

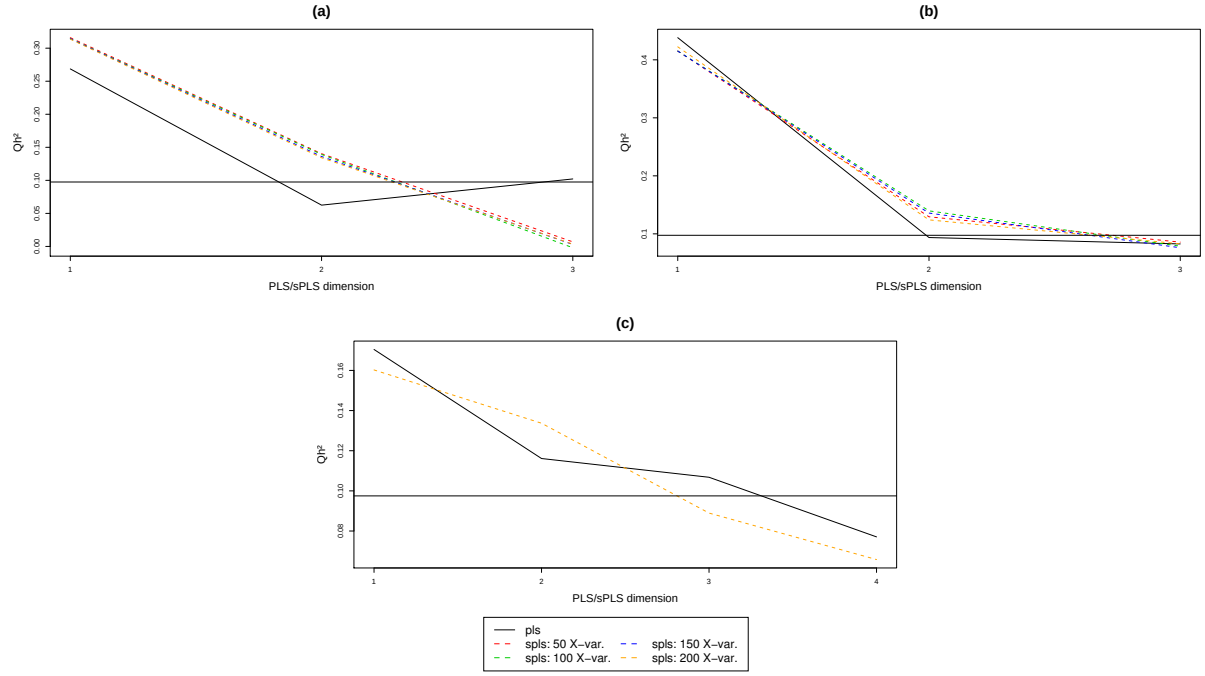

Figure 2: Marginal contribution of the latent variable ξ_h for each component in PLS and PLS (regression mode) with different sparsity degrees for Liver Toxicity Study **(a)**, Wine Yeast **(b)** and Arabidopsis **(c)**. Horizontal black line indicates the threshold value in Q^2_h.

that as the number of samples is very small here (18), the loo-cv computations could only be performed with a sparsity of approx. 200 selected genes in sPLS: removing only one sample in the data set rendered the lasso computation unfeasible (as the original lasso penalizations became too large for the training set).

4.3 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). These graphics show that except for 2 clinical variables, sPLS clearly outperforms PLS. It seems that removing some of the noisy variables in the X data set really seems to help for a better prediction of most of the Y variables. Another extremely interesting fact is that in this figure, clinical variables are ranked according to the absolute value of their loadings in v_2. Hence the Y-loadings really do get a meaning as the less better 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 or not in the biological study.

If the clinical variables were ranked according to the next loading 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 comfort the choice of $H = 2$ for Liver Toxicity with sPLS. Similar conclusions can be drawn on the other data sets that includes more Y variables.

Note also that these kind of graphics should help the user in choosing the penalization parameters as we can observe an improvement from PLS to sPLS in terms of prediction. For

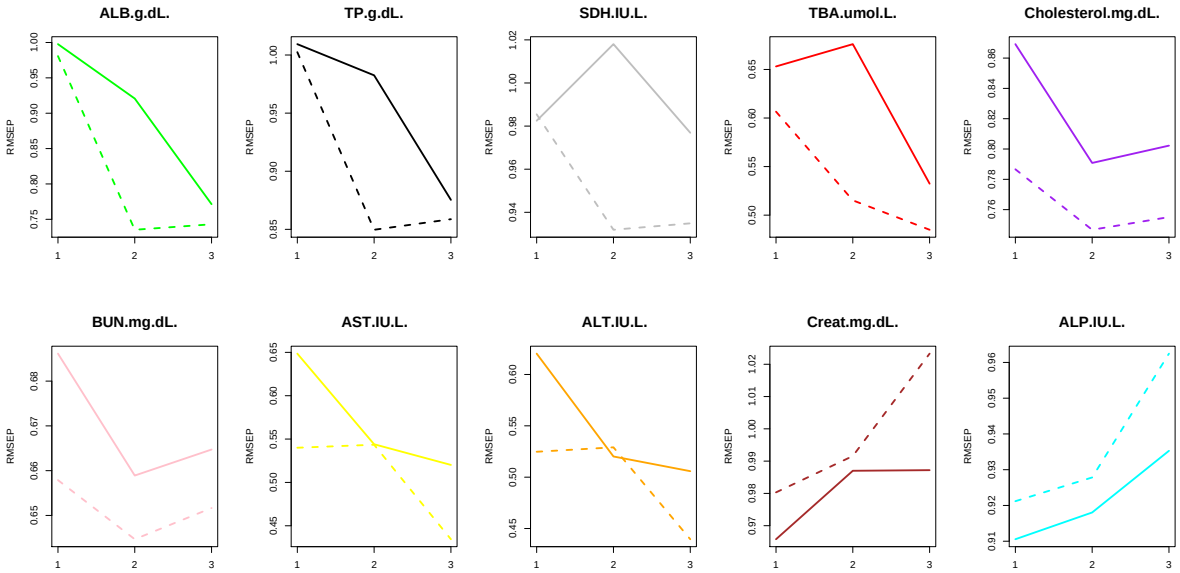

Figure 3: Liver Toxicity study: RMSEP for each clinical variable with PLS (plain line) and sPLS (dashed). Clinical variables are ranked according to their loadings in dimension 2.

example a selection from 50-200 genes gave similar results in Liver Toxicity, but a selection from 150 to 200 genes seemed to improve the prediction in Yeast.

4.4 Graphical representation of the biological samples

Graphical outputs tremendeously help the biologist interpretating the results. In addition, it provides the statistician a tool for assessing the relevance of the tested approach.

Figure 4 displays the latent vectors (ξ_1, ξ_2) and (ξ_2, ξ_3) associated to the cDNA data set in the NCI60 study with PLS (top) or sPLS (bottom) (the other latent variables $(\omega_1, \omega_2, \omega_3)$ showed similar patterns). In both approaches, most of the cancers classes seemed more or less well grouped together, except for Breast where two samples were mixed either with Melanoma or Colon. This was also observed in the studies of Ross et al. (2000); Culhane et al. (2003) where these two cell lines BREAST_MDAN and BREAST_MDAMB435 were found to be melanoma melastases from a patient diagnosed with breast cancer. Remember that there was not much variance explained in any of the PLS, which explains the difficulty to well separate each cancer class.

In PLS (**(a)** and **(c)**), each graphic tended to separate Melanoma and Leukemia cancer, and hence give similar information in any of the three dimensions. In sPLS however, it is clear that dimension 1 clustered the Melanoma together, and the dimension 3 Leukemia and Colon. It would hence be interesting to have a closer look at the genes selection in each dimension to see if each list is closely related to either Melanoma or Leukemia and Colon.

In general these graphics show that sPLS is coherent with the biological study and its latent variables are similar to PLS. In fact, the variable selection performed in sPLS does not seem to affect much the latent variables. These results show that the information about the

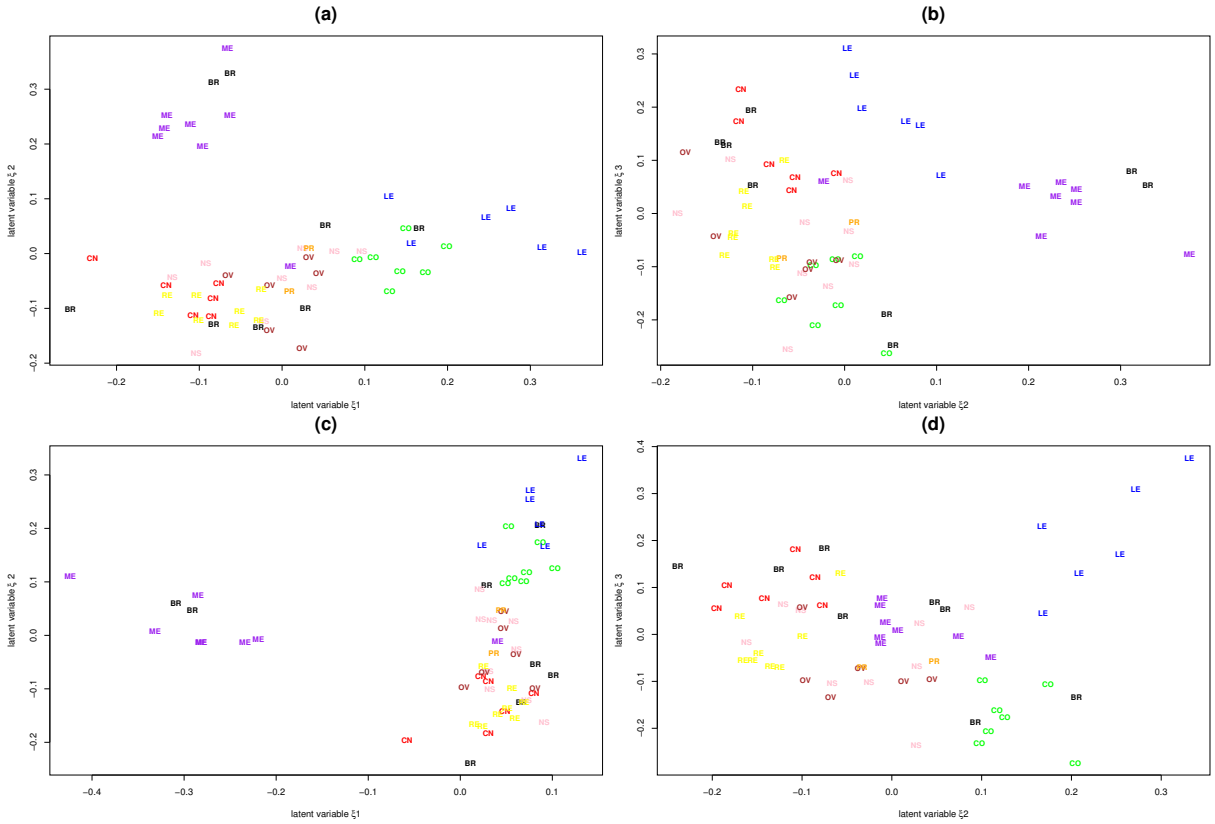

Figure 4: Graphical representation of the samples of the NCI60 study for PLS (top) and sPLS (bottom): cancer classes for the latent vectors (ξ_1, ξ_2) **(a)**, **(c)** and (ξ_2, ξ_3) **(b)**, **(d)**. Cell lines are coded as CO = Colon, ME = Melanoma, BR = Breast, NS= Central Nervous System , OV= Ovarian, RN = Renal, PR = Prostate.

samples remains practically the same, if not better explained when $H > 2$.

4.5 Other interesting facts

The Arapidopsis data set described in Gibon et al. (2006) consisted in only 82 transcripts, selected as coding known enzymes and metboalites from the Y data set. In the sPLS selections that we tested (where 50, 100, 150 and 200 transcripts where selected in each PLS dimension), we found none of these 82 transcripts in dimension 1. In dimension 2 and 3 however appeared 3 to 5 of these transcripts (depending on the selection size). This is not much, but in comparison, a selection performed in two steps with PLS (by first ordering the absolute values of the X-loadings and then selecting from 50 to 200 transcripts artificially) brought no known transcript at all.

It is possible that the transcripts selected with sPLS are linked to unknown metabolites, or that most of the functions of these transcripts are unknown yet, or that these two types of data are not biologically directly related. However these preliminary results may indicate potentially interesting results that would suggest more experiments or further research to be performed.

When applying sparse methods, the loadings may loose their property of orthogonality and uncorrelation, as it was observed with sparse PCA (Trendafilov and Jolliffe, 2006; Shen and Huang, 2007). This is not the case with sPLS. In the original PLS2 (regression mode), 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 in regression mode. To remedy to this in terms of graphical representation of the samples, we reprojected $(\omega_1, \ldots, \omega_H)$ in an orthogonal basis. In the canonical mode however, we always observed that $\xi_r'\xi_s = 0$ and $\omega_r'\omega_s = 0$, $r < s$, meaning that no reprojection is needed at all.

4.6 Biological interpretation of the wine yeast data set

We will first rapidly discuss why the canonical deflation mode is not adapted in this case. We will then give some elements of discussion concerning the graphical representation of the latent variables (samples). These preliminary remarks will explain some of the results obtained when we compared the genes artificially selected with a basic PLS (two-step 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, approx. 200 genes were selected for each dimension.

4.6.1 sPLS modes

The canonical mode did not reveal much in the way of biologically relevant information. This was not entirely unexpected considering that the direction of influence in a living cell is generally from gene to compound. As such the canonical mode (which relies on a reciproqual relationship assumption between both types of data) was in this case both biologically incoherent and impractical.

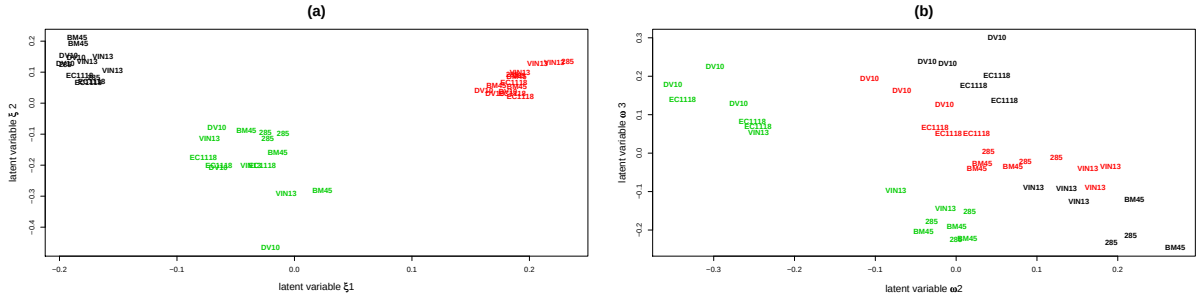

Figure 5: Wine Yeast data : graphical representation of the samples for the latent vectors (ξ_1, ξ_2) **(a)** and (ω_2, ω_3) **(b)**. Colors red, green and black stand for fermentation day 2, 5 and 14.

Table 2: Comparison of genes selected with PLS (two step procedure) vs. sPLS for the regression mode.

	PLS-regression mode	sPLS-regression mode
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

4.6.2 Biological samples

Figure 5 highlights several facts that can actually be explained by the biological experiment. The plots of (ξ_1, ξ_2) (top) and (ω_1, ω_2) gave similar representation (not shown). 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 time (*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.

4.6.3 Selected variables

Comparisons with PLS Table 2 presents the similarities and main differences observed between the genes selected either with PLS or sPLS in regression mode. We adopted a two-

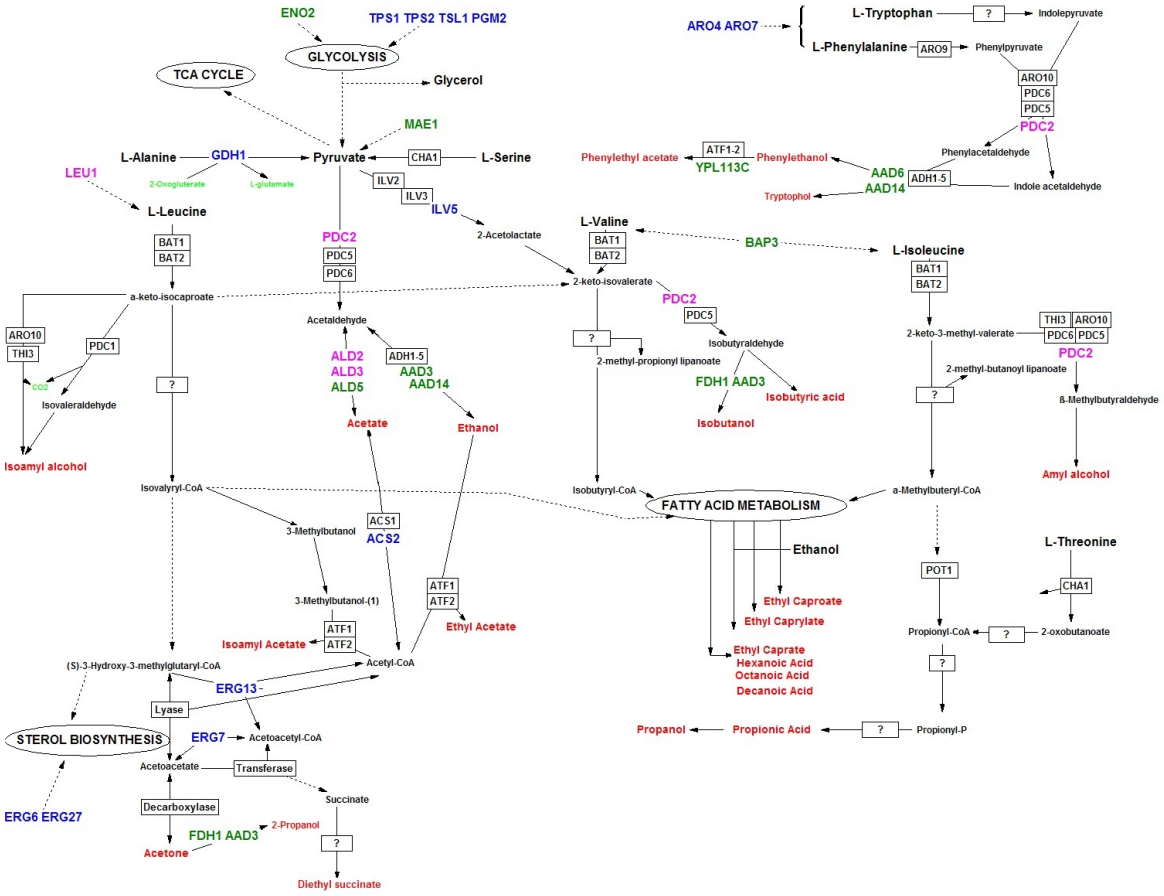

Figure 6: Graphical representation of 'known' or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. Indirect interactions (*i.e* missing intermediates) are indicated by dashed lines and standard reactions are indicated by solid lines. 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.

step procedure to select genes with the original PLS approach by ordering the absolute values of the loadings u_h for each dimension ($H = 3$) and selecting the same number of top genes as in sPLS.

The striking result that we observed was the differences in the genes selections, especially in dimension 2 and 3. Overall, these dimensions were found to be more enriched for genes with proven 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 6 depicts the 'known' or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. 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 4.6.2. 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.

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 ($\xi_1, \dots \xi_H$) and ($\omega_1, \dots \omega_H$), as it is done with PCA or CCA. These graphical representations where the selected variables are projected on a correlation circle, will allow to identify known and unknown relationships between the X-variables, the Y-variables, and more importantly between both types of omics data. Of course these relationship will then need to be biologically assessed with further experiments, and will constitute a next step of our proposed analysis.

5 Conclusion

We have introduced a general computational methodology that modifies PLS, a well known approach that has been proven to be extremely efficient in many data where $n << p + q$, in a sparse version including variable selection to be more useful to the biologists. Two deflation steps are proposed to handle either a regression or a canonical framework of the biological study. Validation of the sparse PLS approach has been performed using four public data sets. Comparisons with the original PLS approach show that sparse PLS benefits not only from the stability property of PLS, but also provides variable selection, an attractive property that facilitate the interpretation of the results.

Like any sparse multivariate method, sPLS requires to add two penalization parameters to choose. However, the algorithm that we propose is very fast to compute, which makes the tuning part easy to perform. The gain by penalizing, and hence selecting variables, is proved on a typical biological study aiming at integrating gene expressions 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. The two deflation modes are also discussed and show that the deflation has to be chosen according to the biological context, in order to get relevant variable selections. This preliminary work undoubtedly brought more insight into the biological study and will suggest further experiments to be performed.

Integrating omics data is a real problem that may soon be commonly encountered in most

high throughput biological studies. Hence we believe that our sparse PLS provides an extremely useful tool for the biologist in need of integrating two-block data sets and easily interpreting the resulting variable selections.

Availability The code sources of sparse PLS (in R[1]) 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. Theor Biol Med Model, 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.

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.

Datta, S. (2001). Exploring relationships in gene expressions: A partial least squares approach. Gene Expr, 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.

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

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

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.

González, I., Déjean, S., Martin, P. G. P., and Baccini, A. (2008). Cca: An r package to extend canonical correlation analysis. Journal of Statistical Software, 23(12).

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.

Hein, A., Richardson, S., Causton, H., Ambler, G., and Green, P. (2005). BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. Biostatistics, 6(3):349.

[1]The Comprehensive R Archive Network, http://cran.r-project.org/

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.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28:321–377.

Johansson, D., Lindgren, P., and Berglund, A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics, 19(4):467–473.

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. Journal of Computational & Graphical Statistics, 12(3):531–547.

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(:Iss. 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).

Nguyen, D. and Rocke, D. (2004). On partial least squares dimension reduction for microarray-based classification: a simulation study. Computational Statistics and Data Analysis, 46(3):407–425.

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

Nykter, M., Aho, T., Ahdesmaki, M., Ruusuvuori, P., Lehmussola, A., and Yli-Harja, O. (2006). Simulation of microarray data with realistic characteristics. BMC Bioinformatics, 7(1):349.

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., Donche, B., and Lonvaud, A. (2000). Biochemistry of alcoholic fermentation and metabolic pathways of wine yeasts in Handbook of Enology, volume 1. John Wiley and Sons.

Ross, D., Scherf, U., Eisen, M., Perou, C., Rees, C., Spellman, P., Iyer, V., Jeffrey, S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 24(3):227–35.

Scherf, U., Ross, D., Waltham, M., Smith, L., Lee, J., Tanabe, L., Kohn, K., Reinhold, W., Myers, T., Andrews, D., et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nat Genet, 24(3):236–244.

Shen, H. and Huang, J. Z. (2007). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, to appear.

Staunton, J., Slonim, D., Coller, H., Tamayo, P., Angelo, M., Park, J., Scherf, U., Lee, J., Reinhold, W., Weinstein, J., et al. (2001). Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences, 98(19):10787.

Steinfath, M., Groth, D., Lisec, J., and Selbig, J. (2008). Metabolite profile analysis: from raw data to regression and classification. Physiologia Plantarum, 132(2):150–161.

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

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.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. Journal of Econometrics, 4(2):147–166.

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.

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.