Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
29 May 2009
Number of times this article has been viewed

Giuseppe Palermo[1]
Paolo Piraino[2]
Hans-Dieter Zucht[3]

[1]Digilab BioVision GmbH, Hannover, Germany; [2]Dr Paolo Piraino Statistical Consulting, Rende (CS), Italy; [3]Proteome Sciences R&D GmbH and C. KG, Frankfurt am Main, Germany

Abstract: Multivariate partial least square (PLS) regression allows the modeling of complex biological events, by considering different factors at the same time. It is unaffected by data collinearity, representing a valuable method for modeling high-dimensional biological data (as derived from genomics, proteomics and peptidomics). In presence of multiple responses, it is of particular interest how to appropriately "dissect" the model, to reveal the importance of single attributes with regard to individual responses (for example, variable selection). In this paper, performances of multivariate PLS regression coefficients, in selecting relevant predictors for different responses in omics-type of data, were investigated by means of a receiver operating characteristic (ROC) analysis. For this purpose, simulated data, mimicking the covariance structures of microarray and liquid chromatography mass spectrometric data, were used to generate matrices of predictors and responses. The relevant predictors were set a priori. The influences of noise, the source of data with different covariance structure and the size of relevant predictors were investigated. Results demonstrate the applicability of PLS regression coefficients in selecting variables for each response of a multivariate PLS, in omics-type of data. Comparisons with other feature selection methods, such as variable importance in the projection scores, principal component regression, and least absolute shrinkage and selection operator regression were also provided.

Keywords: partial least square regression, regression coefficients, variable selection, biomarker discovery, omics-data

Introduction

The analysis of high dimensional biological data, as derived from omics-type data (for example, genomics, proteomics, and peptidomics) is a very challenging task. A limited amount of samples with thousands of features, give rise to known issues, as data overfitting and multicollinearity. Moreover, the complex pattern of biological events can depend on different factors that must be included in the analysis for a proper description of the model. Multivariate partial least square (PLS) regression allows the modeling of multiple responses, while dealing with multicollinearity.[1] It can be used for variable selection, as a process to discover the most relevant features of the model (these attributes can be used as biomarker candidates).[2] In multivariate PLS, it is of interest to "dissect" the importance of single attributes, with regard to individual responses. It will exploit the holistic model of responses as offered by a multivariate PLS, while focusing onto variables that are important to a specific response. The aim of this paper is to select variables "independently" for each response of a multivariate PLS. A recent work has compared the performance of the so-called variable importance in the projection (VIP) scores[3] with PLS regression coefficients, to select variables for

Correspondence: Hans-Dieter Zucht
Proteome Sciences R&D GmbH and Co. KG, Altenhöferallee 3, D-60438, Frankfurt am Main, Germany
Email hans-dieter.zucht@proteomics.com

single-response PLS models.[4] They have considered the case with more observations than features ($n > p$). Another work has studied variable selection for the case $n << p$, based on single response PLS.[5] This paper considered the case p >> n (as it is common for omics-type of data), to select features from each response of a multivariate PLS. In detail, simulated data, mimicking the covariance structure of real microarray and liquid chromatography mass spectrometric (LC-MS) data, were used to investigate the performance of PLS regression coefficients in variable selection. A two-response PLS was first considered, as a model case, further drawing conclusions on a PLS with more responses. In the simulation, responses were generated from true models. Only few predictors were relevant to a response, meaning that they had nonzero regression coefficients. Those relevant predictors were set *a priori*, with the requirement that they were correlated each other. The performance of PLS regression coefficients, in selecting relevant predictors, could then be investigated by means of the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. Results were compared with other methods which can be also used for variable selection, such as principal component regression (PCR), VIP scores and least absolute shrinkage and selection operator (Lasso).

Methods

PLS

The PLS model is based on principal components on both the independent data, X, and the dependent data, Y. The basic idea is to calculate the principal scores of X and Y and to set up a regression model between the scores.

$$\begin{aligned} X &= T\,P' + E \\ Y &= U\,Q' + F \end{aligned} \tag{1}$$

Thus the matrix, X, is decomposed into a matrix, T (referred to as X-score), and a matrix, P' (referred to as X-loading), plus an error matrix, E. The matrix, Y, is decomposed, equivalently, into the Y-scores, U, the Y-loadings, Q', and the error term, F. These two equations (1) are called outer relations, and they model X and Y respectively by the score vectors T and U. The goal of the PLS algorithm is, then, to minimize the norm of F while keeping the correlation between X and Y by the inner relation $U = TD$, where D is a diagonal matrix. The X-scores are orthogonal. They are estimated as linear combinations of the original variables x_k with the coefficients, *weights* w^*_{kl} ($k = 1, 2, \ldots, p$; $l = 1, 2, \ldots, a$ where a is the number of components in the model).

$$T = XW^* \tag{2}$$

PLS, then, can be seen as a method to construct a matrix of latent variables as a linear transformation of X, where $W^*(p \times a)$ is a matrix of weights.

Using the inner relation,

$$\begin{aligned} Y &= UQ' + F = TDQ' + (HQ' + F) = TC' + F^* = \\ &= XW^*C' + F^* = XB + F^* \end{aligned} \tag{3}$$

with B $(p \times m)$, referred to as PLS regression coefficients, equal to

$$B = W^* * C' \tag{4}$$

Different numeric algorithms, to obtain a solution of the PLS regression problem, appear in the literature. For instance, the nonlinear iterative partial least squares (NIPALS) algorithm can be used to sequentially extract the PLS components; details on the NIPALS algorithm can be found in.[6]

PLS regression coefficients can be used to select relevant predictors according to the magnitude of their absolute values.[4]

An alternative method for variable selection based on PLS regression is the so-called VIP, first published in.[7] The VIP score of a predictor is a summary of the importance for the projections to find a latent variables. VIP values can be calculated by summing variable influence (VIN) over all model dimensions.[2] For a given PLS dimension a, $(VIN)_{ak}^{\ 2}$ is equal to the squared PLS weight $(w_{ak})^2$ of that term, multiplied by the percent explained of residual sum of squares by that PLS dimension. The accumulated (overall PLS dimensions) value, $VIP_k = \sum_a \left((VIN)_{ak}^{\ 2} \right)$, is then divided by the total percent explained of residual sum of squares by the PLS model and multiplied by the number of terms in the model. VIP scores can be used to select relevant predictors according to the magnitude of their values.[4]

PCR

Principal component regression (PCR) is a two-step multivariate calibration method. In the first step a principal component analysis (PCA) of the matrix, X, is performed. The measured variables are converted into new ones (scores and latent variables). This is followed by a multiple linear regression step (MLR) between the scores obtained in the PCA step and the response matrix, Y.

PCA creates new orthogonal variables (latent variables) that are linear combinations of the original x-variables.

$$X = TP' \tag{5}$$

T is the score matrix. P is the loading matrix. Two main advantages arise from this decomposition. The first one is

that the new variables are orthogonal. Then the inversion of T (needed in the MLE step) is no longer a problem, as it is when original variables are correlated. Moreover, it is assumed that the first few PCs, accounting for the majority of the variance of the original data, contain meaningful information, while the last ones can be deleted. Therefore only r < min(n, p) PCs are retained, obtaining a simplified model. After performing PCA on X, the second step in PCR consist of the linear regression between the scores and the response matrix Y, which is modeled by

$$Y = TC + E = XPC + E = XB + E \quad (6)$$

with the regression coefficients given by

$$B = P\,(T'T)^{-1}\,T'Y \quad (7)$$

Least absolute shrinkage and selection operator

The Lasso is a shrinkage and selection method for linear regression. It is a constrained version of ordinary least squares. It minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant s. If the data are standardized to have mean 0, the Lasso estimate is defined by equation (8). The tuning parameter, $s \geq 0$, can be determined by cross validation. Because of the nature of the constraint, it tends to produce some coefficients as zero and it may improve the overall prediction accuracy by sacrificing a little bias to reduce the variance of the predicted values.

$$\hat{\beta}^{lasso} = \arg\min_{\beta}(y - X\beta)\,(y - X\beta) \text{ subject to } \sum_{j=1}^{p}|\beta| \leq s \quad (8)$$

In this work, the Lasso regression coefficients were calculated with the least angle regression (LARS) method[8] implemented in the *R*′s package *LARS*.[9] Details of the LARS algorithm for the Lasso estimate can be found in Chong and Jun.[4]

ROC curve as performance measure in selecting relevant predictors

In order to use the multivariate PLS regression coefficients to find relevant predictors, the corresponding density distributions for relevant and irrelevant predictors should only moderately overlap (see Figure 1). The task of finding

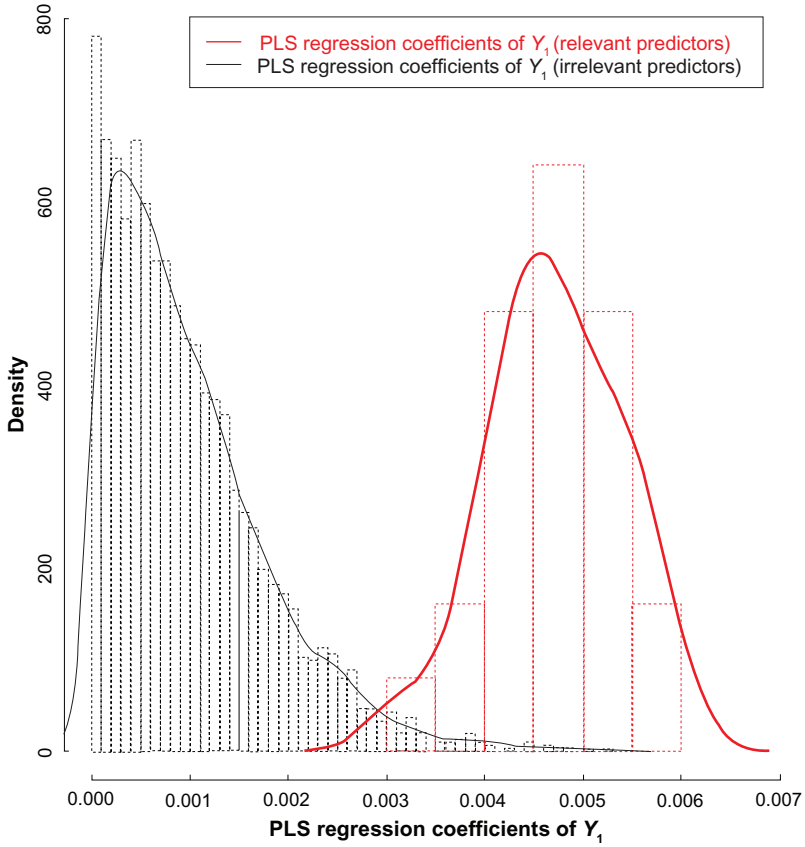


Figure 1 Density distributions of the absolute values of multivariate partial least square (PLS) regression coefficients for (left) irrelevant and (right) relevant predictors. A multivariate PLS was used to model a response matrix $Y = (Y_1, Y_2)$, with 100 observations. The matrix of predictors X was simulated from a real microarray dataset. The size of predictors was 3751. The relevant predictors (1% of total predictors) were known *a priori*.

relevant predictors, for a given response, can be seen as a two-class discrimination problem. The two classes, in this case, refer to relevant and irrelevant predictors (in the following of this section also referred as to positive and negative classes). Sensitivity and specificity are the basic measures of accuracy for a classification task. They can be obtained from the confusion matrix (Table 1), which contains information about actual and predicted classifications done by a classification system.

Sensitivity is a statistical measure of how well a binary classification test correctly identifies a condition (positive class; relevant predictors). It represents the proportion of true positive cases of all positive cases in the population. Specificity represents the proportion of true negative cases of all negative cases in the population. Using the notation from Table 1,

$$\text{Sensitivity} = a/(a+b)$$
$$\text{Specificity} = d/(c+d)$$
$$\text{False positive rate} = 1 - \text{specificity}$$

where the false positive rate (FPR) represents the proportion of actual negative cases wrongly assigned to the positive class.

The ROC is a plot of sensitivity versus its false-positive rate (FPR) for all possible cut points, illustrating how sensitivity and FPR vary together.[10,11] One of the most decisive measure of accuracy for a classification test is then the area under the ROC curve (ROC-AUC).[11] The practical range for the ROC-AUC is between 0.5 and 1.0. A test with a ROC-AUC of 1.0 is perfectly accurate, because the sensitivity is 1.0 and the FPR is 0.0 (meaning that all relevant predictors were correctly identified, without irrelevant predictors wrongly assigned to the positive class). In contrast, a value of 0.5 corresponds to a test that is purely guessing the result (the probability to detect a truly relevant predictor, in this case, is equal to a flip of coin). The ROC-AUC can be interpreted as the average value of sensitivity for all possible values of specificity.

Experimental design

Design of simulation

Simulated data were used to investigate the performance of PLS regression coefficients, to select relevant predictors independently for each response of a two-response PLS. For this purpose, datasets were generated by assuming a linear relationship between true responses Y and the matrix of predictors X, as defined by

$$Y = (Y_1, Y_2) = XB + Error = X \cdot (\alpha, \beta) + (\varepsilon, \delta) \quad (9)$$

$Y_1 = \left(y_1^1, ..., y_n^1\right)^t$ and $Y_2 = \left(y_1^2, ..., y_n^2\right)^t$ are the true response vectors. The number of observations, n, was arbitrarily fixed to 100, being a reasonable choice, given the number of samples usually employed in omics-type studies. $X = x_{ij}$ ($i = 1, ..., n$; $j = 1, ..., p$) is the matrix of predictors (p is the total size of predictors). It was generated using the covariance structure of real datasets. For this purpose, three microarray datasets were considered. In addition, an unpublished tab delimited LC-MS dataset was used. $\alpha = (\alpha_1, ..., \alpha_p)^t$ and $\beta = (\beta_1, ..., \beta_p)^t$, in (9), are regression coefficients, respectively, for Y_1 and Y_2. Regression coefficients corresponding to relevant (irrelevant) predictors were set to 1.0 (0.0). The size of relevant predictors was set to a fixed percentage of the total number of predictors, p. $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^t$ and $\delta = (\delta_1, \delta_2, ..., \delta_n)^t$, in (9), are the error terms, respectively, for Y_1 and Y_2. They were distributed according a standard distribution $(\varepsilon_j \approx N(0, \sigma_1^2), (\delta_j \approx N(0, \sigma_2^2), i = 1, 2, ..., n)$. In summary, an experimental design with 36 (= 4 × 3 × 3) different cases and three factors was considered: the real dataset from which X was generated (4 levels), the proportion of relevant predictors among all predictors (3 levels) and the magnitude of signal to noise (3 levels). In each case 100 replications were made. At each replication, a different dataset of 100 observations was generated according to equation (9). A PLS model was then calculated. Finally, the performance of multivariate PLS regression coefficients, in selecting relevant predictors, was calculated by means of a ROC analysis. Details on factors that were considered in the experimental design are provided in the next sections.

Table 1 Confusion matrix

		Predicted	
		Positive	**Negative**
Actual	Positive	a	b
	Negative	c	d

Factor 1: The influence of the dataset used in the simulation

Four real datasets (see Table 2) were used in the simulation. The *leukemia* dataset[12] has frequently been used in previous microarray data analysis studies. It contains the expression levels of 7129 genes for 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) patients. Data were preprocessed following the procedure described in,[13] remaining with 3751 variables.

The *colon* dataset[14] is an other benchmark dataset, frequently used for testing different methods on gene expression data.

Table 2 Real datasets used to simulate a matrix *X* of predictors

Dataset	*n*	*p*
Colon	62	2,000
Leukemia	72	3,571
SRBCT	63	2,308
Alzheimer	92	2041

Abbreviation: SRBCT, small round blue cells tumor.

It consists of the expressions for 6500 genes, measured on 62 samples: 22 healthy patients and 40 colon cancers. 2000 genes were selected by the authors for clustering/classification purpose.

The *SRBCT* dataset[15] consists of the expression for 2308 genes, measured on 83 samples from small round blue cells tumor (SRBCT), belonging to four subclasses: non-Hodgkin lymphoma (BL), Ewing family of tumors (EWS), rhabdomyosarcoma (RMS) and neuroblastoma (NB).

Finally, the *Alzheimer* dataset[16] consists of spectrometric data, where cerebrospinal fluid (CSF) of Alzheimer disease (AD) and nondemented controls were compared, to find peptides likely to correlate with the AD pathogenesis. The dataset included 2041 signals measured on 45 AD samples and 47 controls. Profiling of peptides was based on MALDI mass-spectrometric analysis of samples, previously fractionated by reverse-phase chromatography, to reduce their complexity.

Leukemia, Colon and SRBCT datasets were all available in the *R*´s package *plsgenomics*.[17,18] The Alzheimer dataset was unpublished. The matrix of predictors *X*, in equation (9), was generated mimicking the covariance structure of datasets from Table 2. The number of samples, in *X*, was fixed to 100, as explained in the Design of simulation section. The number of predictors was equal to $p = 2.000$, $p = 3.571$, $p = 2{,}308$ or $p = 2.041$, depending on the source of simulation (Table 2). Details on the algorithm that was used to simulate *X* can be found in the supplementary material.

Factor 2: The influence of size of relevant predictors

The percentage of relevant predictors, among all predictors, was arbitrarily set to one of the following three levels:

$$\text{Levels for the percentage of relevant predictors} = 1\%, 2\%, 5\% \quad (10)$$

This is equivalent to the assumption that only a small percentage of variables are relevant to a response. The relevant predictors were chosen to be correlated each other. In microarrays studies it has already been shown that clustering gene expression data groups together related genes.[19] Then, the hypothesis that a cluster of genes may be relevant to model a phenomena *Y* is plausible. In order to group predictors with a similar profile, an unsupervised hierarchical clustering algorithm was applied to the matrix *X* (the Pearson's correlation coefficient was chosen as similarity measure). Two branches of the cluster, C_1 and C_2, were randomly selected (for example see Figure 2). Their size was chosen according to equation (10). Predictors belonging to C_1 and C_2 were set as relevant predictors, respectively, for Y_1 and Y_2. Mathematically, it can be obtained as

$$\begin{aligned} &\alpha_i = 0.0 \text{ if predictor}_i \notin C_1 \quad \beta_i = 0.0 \text{ if predictor}_i \notin C_2 \\ &\alpha_i = 1.0 \text{ if predictor}_i \notin C_1 \quad \beta_i = 0.0 \text{ if predictor}_i \notin C_2 \end{aligned} \quad (11)$$

for $i = 1, 2, \ldots, p$. $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^t$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^t$ are the regression coefficients in equation (9). An equal contribution to responses, from relevant predictors, was considered.

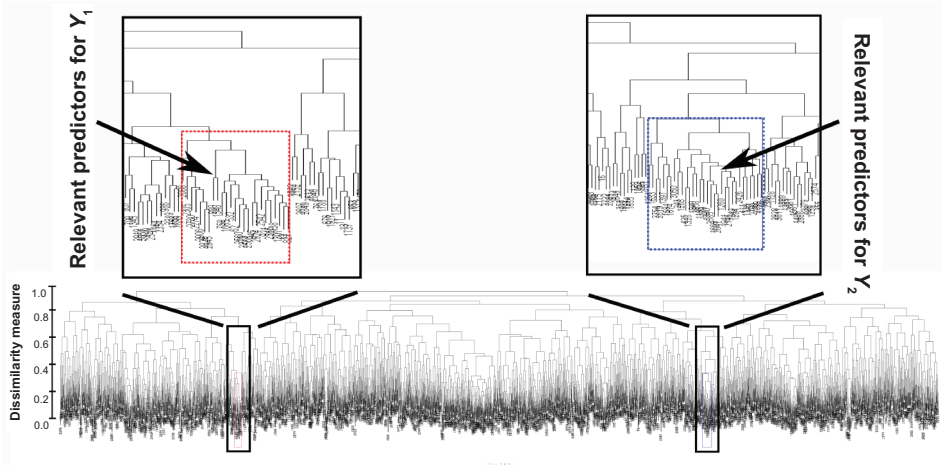


Figure 2 An hierarchical clustering was performed on the matrix of predictors *X*, simulated using the covariance structure of the leukemia dataset. Two branches of the cluster, colored in red and blue, were arbitrarily selected as relevant predictors, respectively, for Y_1 and Y_2.

Consequently, regression coefficients of Y_1 (Y_2), for predictors belonging to C_1 (C_2), were set to 1.0. All regression coefficients corresponding to irrelevant predictors were set to 0.0.

Factor 3: The influence of the magnitude of noise

In this paper, an important issue was to investigate how noise, in equation (9), will affect performance of PLS regression coefficients in variable selection. Following a recent work,[4] three levels for the error terms in equation (9) were considered, defined by

$$\sigma_1 = k\sqrt{\text{var}(X\alpha)}, \sigma_2 = k\sqrt{\text{var}(X\beta)} \quad (12)$$

with k, the reciprocal of the signal to noise ratio, equal to $k = 0.33$, 0.74, 1.22, and var(·) in equation (12) representing the sample variance. These levels were chosen such that R-square of the multiple linear regression with an intercept become 0.9, 0.65 and 0.4, respectively, when infinite observations are assumed.[4] Some simple calculations using the formula for R-square were also given: $k = ((1-R^2)/R^2)$

The response matrix *Y*

Once a matrix of predictors, X, was simulated (as explained in the section on Factor 1) and the relevant predictors for each response were set by defining the regression coefficients B (through a cluster analysis of X, as explained in the section on Factor 2), the response matrix Y could be generated according to equation (9).

Results and discussion

Following the experimental procedure described in Experimental design, 100 replications for each of 36 cases were considered, to evaluate the performance of PLS regression coefficients in selecting variables independently for each response of a two-response PLS. A PLS regression model was fitted for each case and each replication, using a 10-fold cross validation as a criteria to choose the number of latent variables (PLS components) in the model. The NIPALS algorithm for PLS regression was used through the all study.

Figure 3 shows ROC plots for all 36 cases, providing a performance measure for all conditions in the simulation. Each curve represents an average ROC curve on the responses Y_1 and Y_2 (the average was calculated on 200 ROC curves: 100 replications for each response). Correspondingly, Figure 4 plots the ROC-AUC values for all cases, redundantly, to provide a description for the main effects and interactions of factors in the simulation schema.

Results from Figure 3 and Figure 4 show that performance of variable selection, based on PLS regression coefficients, is robust against noise (increasing k, in the section on Factor 3, from 0.33 to 1.22, in average, decreases the ROC-AUC of 3.1%). In contrast, performance is significantly affected by the size of relevant predictors (increasing size from 1% to 5%, in average, decreases the ROC-AUC of 8.5%). Results further suggest a significant interaction between noise and size of relevant predictors (increasing k from 0.33 to 1.22 decreases, in average, the ROC-AUC of 2.2% and 4.7 %, depending on the size of relevant predictors being equal or different than 5%).

A reason why the size of relevant predictors affects variable selection performance, is related to the fact that the overall correlation between relevant predictors is dependent on their size, due to current experimental design. In fact, since relevant predictors were set as belonging to a branch of a cluster (see the section on Factor 2), in order to increase their size, it is required to choose a bigger branch. This can be obtained by selecting a new node in the dendrogram (see Figure 5), at an higher level of dissimilarity, which in turn weakens the overall correlation between the increased number of predictors. As a consequence, sensitivity/specificity of variable selection is decreased.

Some evidences that performance of PLS regression coefficients, in variable selection, is strongly dependent on the correlation between relevant predictors, was given by means of an additional simulation. In detail, the Colon dataset was used to generate a matrix X of predictors. The noise factor was set to its lowest level ($k = 0.33$, see the section on Factor 3). Two groups of predictors (with size equal to 1% of total size of predictors) were "randomly" chosen as the relevant predictors, respectively, for Y_1 and Y_2. This time, since relevant predictors did not belong to a branch of a cluster, they were not expected to be significantly correlated each other. In this case, a ROC analysis for selection of relevant predictors, using PLS regression coefficients, gave a ROC-AUC value of 0.67 (data not shown), as compared to 0.98, when relevant predictors were grouped into a cluster of X (results were averaged on 100 replications of the above simulation).

Looking at Figure 4, it can be seen that as the size of predictors increases, the negative trend for the AUC is less significant for the Colon dataset, as compared with other datasets from Table 2. In fact, increasing the size of relevant predictors from 1% to 5%, decreases, in average, the ROC-AUC of 4.6%, 10.1%, 13.9%, and 9.4%, respectively, for the Colon, SRBCT, Leukemia, and Alzheimer datasets.

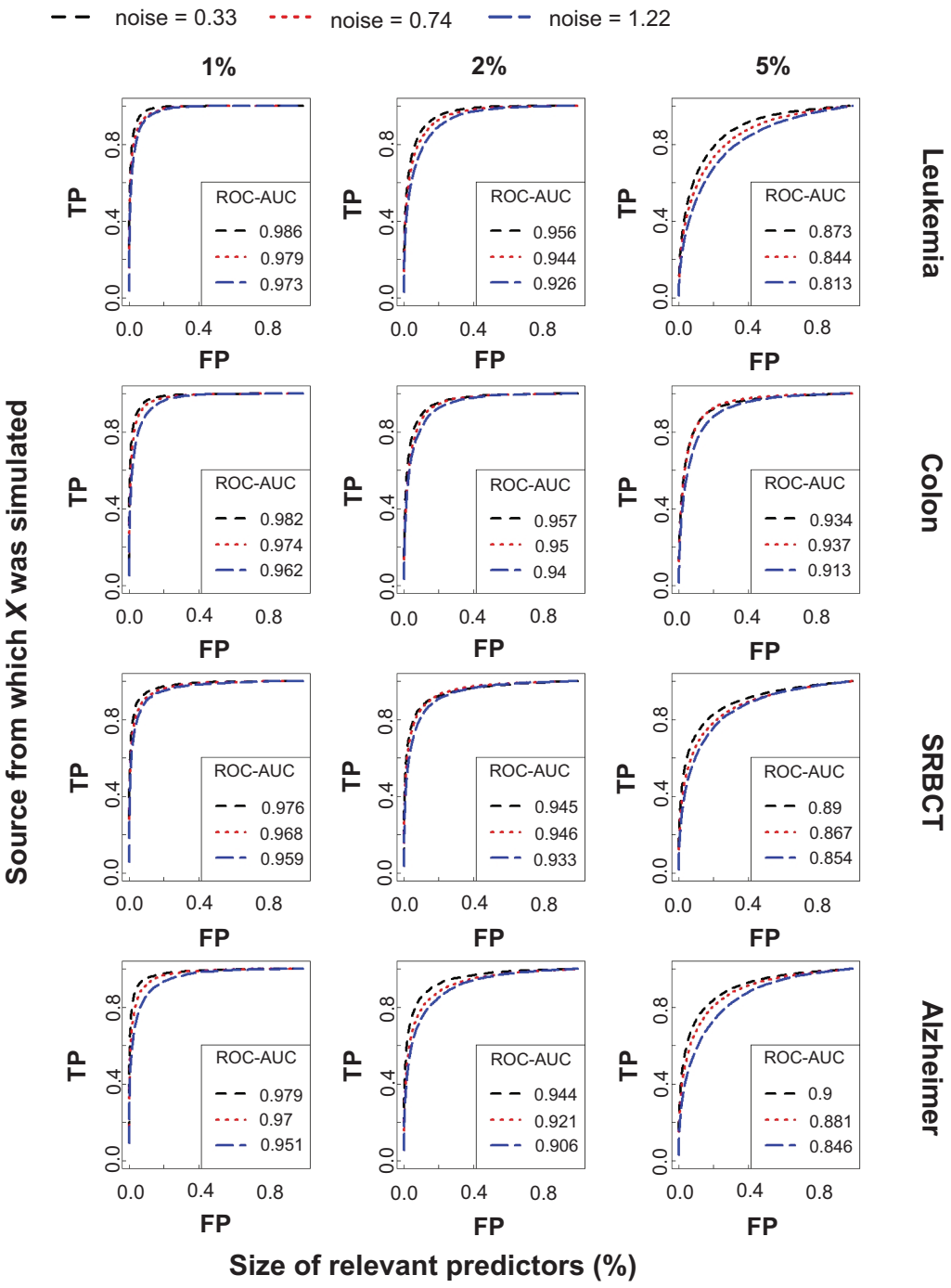


Figure 3 The performance of PLS regression coefficients, in selecting variables independently for Y_1 and Y_2, is assessed by means of a ROC analysis. ROC curves are evaluated for each of 36 cases of the experimental design. Each curve is an average on the two responses Y_1 and Y_2 (the average was calculated on 200 ROC curves: 100 replications for each response).

Abbreviations: AUC, area under the curve; FP, false positive; PLS, partial least square; ROC, receiver operating characteristic; SRBCT, small round blue cells tumor; TP, true positive.

One reason for differences of performance in the four datasets is their different covariance structures. The first 5 components of a principal component analysis explained 71%, 42%, 36%, and 56% of total variance, respectively for the Colon, Leukemia, SRBCT, and Alzheimer datasets. These differences were already visible by comparing the hierarchical clustering of the four datasets (data not shown). For example, to select a node with 5% of predictors in the corresponding dendrograms, a cutoff threshold above 0.6 (in the dissimilarity range 0.0–1.0) was required for Leukemia, SRBCT, and Alzheimer datasets, as compared to a lower threshold of 0.4 for the Colon dataset.

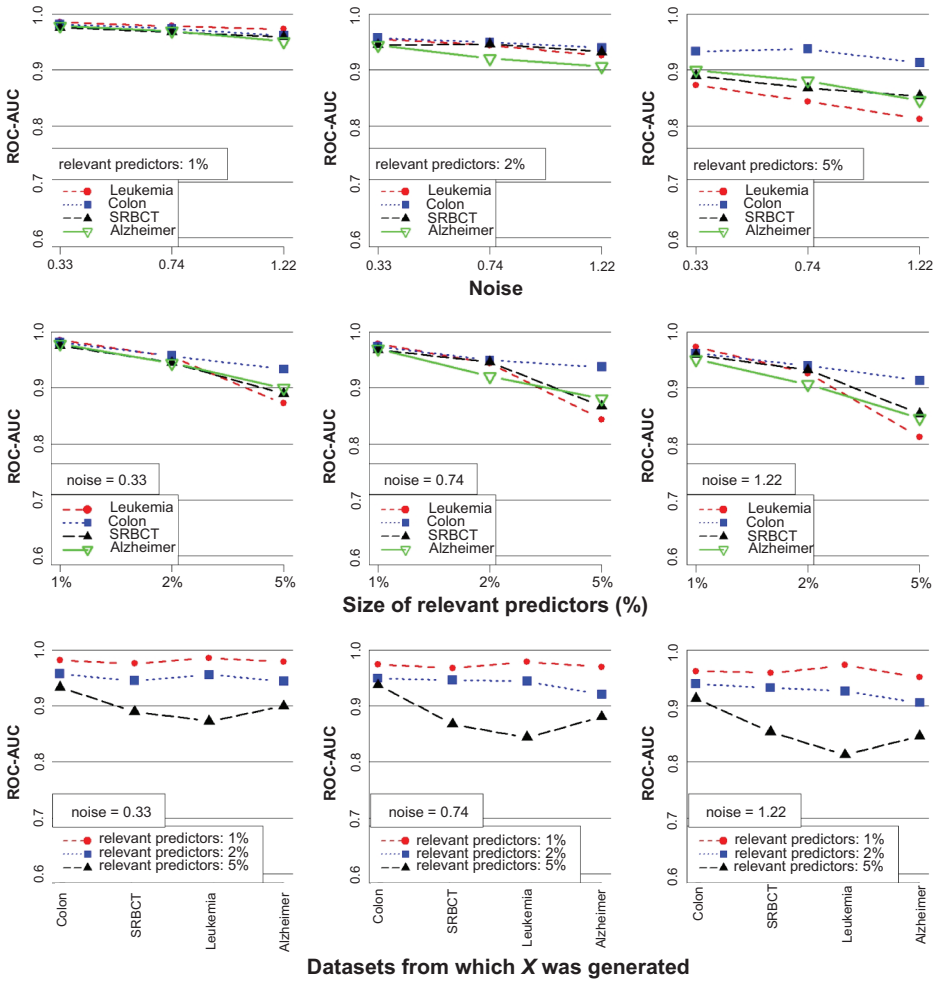


Figure 4 The ROC-AUC values summarize the results of the ROC analysis in selecting variables independently for Y_1 and Y_2. ROC-AUC values are calculated for each of 36 cases of the experimental design, based on the corresponding ROC curves. Each point is an average on the two responses Y_1 and Y_2 (the average was calculated on 200 ROC curves: 100 replications for each response). A redundant representation of 36 averaged ROC-AUC describes the main effects and interactions of factors of the experimental design.

Abbreviations: AUC, area under the curve; FP, false positive; PLS, partial least square; ROC, receiver operating characteristic; SRBCT, small round blue cells tumor; TP, true positive.

Current results are, in general, valid for different levels of correlation between Y_1 and Y_2, since $cor(Y_1, Y_2)$ used to vary across replicated runs in the simulation. Anyway, it was checked if there was a relation between the obtained AUC (in selecting relevant predictors) and the correlation between Y_1 and Y_2. Interestingly it was not found any repetitive pattern for AUC performance with correlation changes between responses.

The overall approach could have been easily applied to a PLS with more than two responses. In this work, the same schema that was used for a two-response PLS, was also adapted to a PLS with three and four responses (selecting respectively three and four branches, instead of two, from the hierarchical clustering, as explained in the section on Factor 2). No significant differences in performance were observed between two-, three- and four-response PLS, when PLS regression coefficients were used for variable selection (results for three- and four-response PLS can be found in the Supplementary material, Figures S1 and S2).

Three other feature selection techniques were considered in the simulation. Specifically, performances of PCR, VIP, scores and Lasso regression in variable selection were compared to PLS regression coefficients for the case with two responses (2-columns *Y* matrix). The same experimental procedure as for PLS was used (36 cases with 100 replications; see Design of simulation). For each replication, a 10 fold cross-validation was used to choose the number of components of the PCR model from which PCR regression coefficients were estimated. For the VIP scores, a univariate PLS regression model was fitted for each response and a 10-fold cross-validation was used to choose the number of components. VIP scores were then calculated from each model as explained in the section on PLS, above. Finally, Lasso regression coefficients were estimated for each response

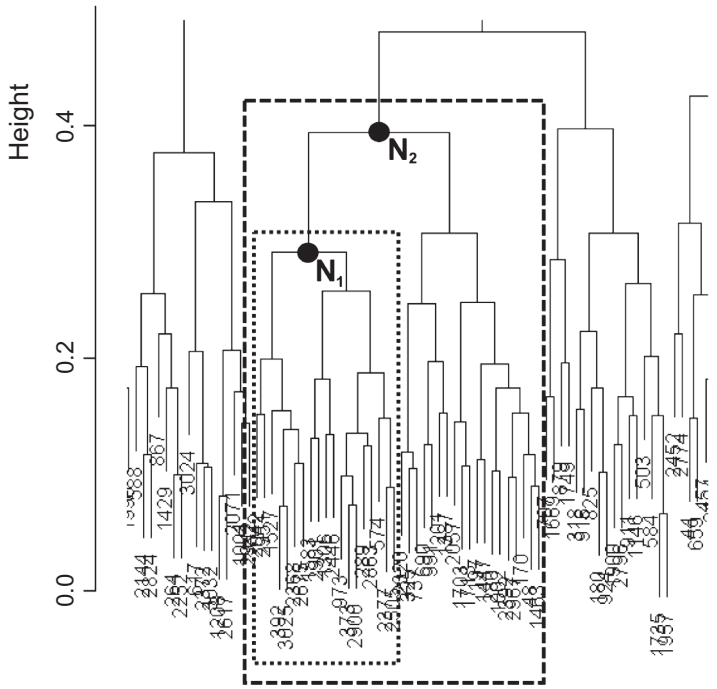


Figure 5 Small window on the hierarchical clustering of the leukemia dataset. Increasing the number of relevant predictors, requires the selection of a new node (for instance N_2), at a higher level of dissimilarity in the y-scale.

according to equation (8) and a 10-fold cross validation was used to estimate the tuning parameter, s. Performances of these methods to select variables were assessed by means of a ROC analysis applied respectively on the absolute values of the PCR regression coefficients, on the absolute values of the Lasso regression coefficients and on the VIP score values.

Figure 6 compares the ROC-AUC values for all the variable selection methods which were considered in this study. Results were summarized by two factors: noise and size of relevant predictors (see sections on Factor 2 and Factor 1, respectively). For each level of those factors, a mean ROC-AUC value was calculated as an average across all the replications considering that level. Results for PLS regression coefficients, PCR regression coefficients and VIP scores were comparable, although VIP scores slightly outperformed both other methods for all cases. All the three methods significantly outperformed Lasso regression coefficients. No significant differences in performance were observed between PLS and PCR regression coefficients. Similar results were found by[4] which compared VIP scores, PLS regression coefficients and Lasso regression coefficients to selected variables for a single-response Y and $p < n$.

Conclusions

In this paper, simulated data, mimicking the covariance structure of real microarray and LC-MS data, were used to explore the performance of PLS regression coefficients in selecting variables independently for each response of a two-response PLS. The response vectors, Y_1 and Y_2 were modeled according true models. It was assumed that relevant predictors were few and correlated each other. It was investigated how variable selection performance, of PLS regression coefficients, was influenced by three factors: the real dataset from which X was simulated, the magnitude of the noise and the size of relevant predictors. The results showed that the method appears relatively robust against the presence of noise. Rather it was dependent on the size of relevant predictors, caused mostly by varying correlation levels between relevant predictors. In fact, since the overall correlation between relevant predictors increases with their size (due to current experimental design, see Discussion), the two effects (correlation and size of relevant predictors) were confounded. However, it was shown that ROC performance decreases drastically in case relevant predictors were not correlated each other. This indicates that presence of correlation between relevant predictors has a big impact on performance of the variable selection strategy. Current results, also, showed that best performances were achieved with the Colon dataset. A deeper analysis of the four datasets unmasked differences in their covariance structures This was based on a principal component analysis, as well on comparisons of their inner dissimilarity representations,

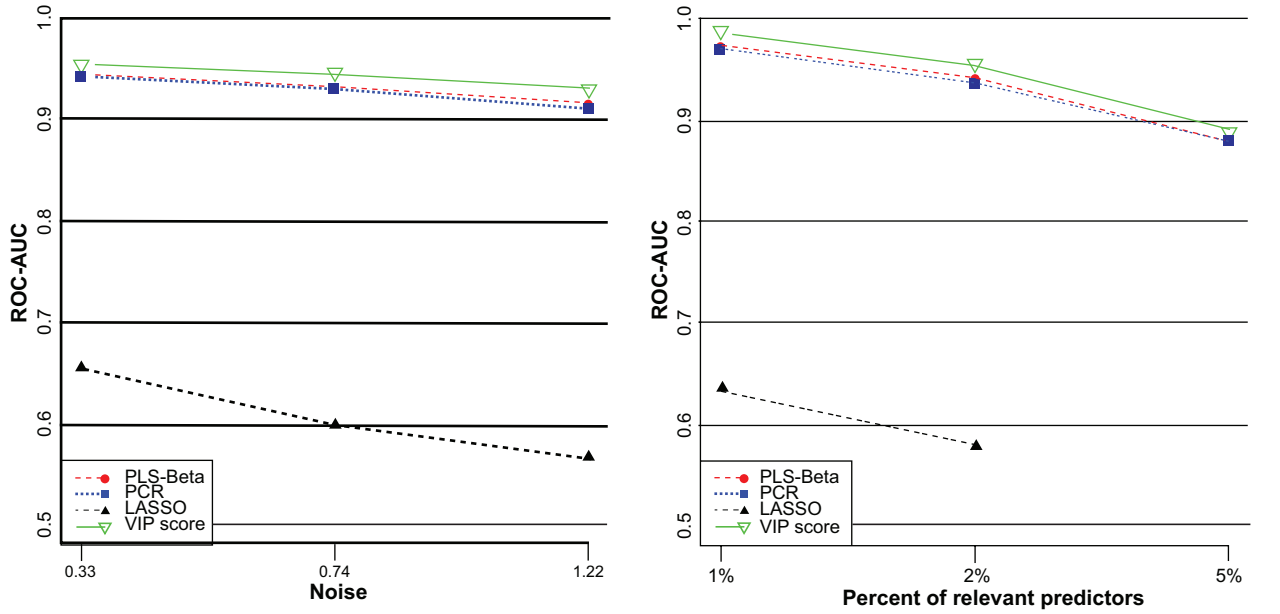


Figure 6 Variable selection performances of PLS regression coefficients (PLS-Beta), PCR coefficients (PCR), Lasso regression coefficients (LASSO) and VIP score were compared. Results were summarized by two factors: noise and size of relevant predictors. Results for LASSO at the 5% level of relevant predictors could not be obtained due to the LASSO implementation based on LARS, which imply no more than *n-1* variables with no-zero coefficients (with *n* the sample size).
Abbreviations: AUC, area under the curve; PCR, polymerase chain reaction; PLS, partial least square; ROC, receiver operating characteristic.

as provided by a cluster analysis. In this respect, the Colon dataset revealed an higher similarity between its variables, as compared to the Leukemia, SRBCT, and Alzheimer datasets. It suggests that better performances are achievable as stronger the predictors are correlated each other. To give some clue that PLS regression coefficients can be used, as well, for selecting variables independently for more than two responses, the simulation schema considered for a two-response PLS, was extended to three- and four-response PLS. Results for three- and four-response PLS were almost identical to the two-response case.

It is, of course, clear that univariate PLS could have been consistently used for modeling each response to select features. In this case, as many univariate PLS models as different responses would have to be calculated. Then, for each model, univariate PLS regression coefficients could be used to extract relevant features for the corresponding response. It is not difficult to believe that the above strategy would bring equivalent results in selecting features as with the multivariate PLS approach (data not shown). However, this means that a multivariate PLS alone can be used in place of *k* univariate PLS regressions (with *k* the size of responses). As a consequence, the output of PLS will be more compact in keeping track of a single model instead of *k* models. A further advantage is that the different responses will be modeled on the basis of the same principal components. Which in turn will allow to exploit relationships between responses, as, for instance, highlighted by a loading-loading plot, where all responses are simultaneously represented.

The number of the PLS components to include in the final model is central and difficult in the PLS regression framework. In the case of univariate PLS applied to binary classification problems, the weight vector $w_1 = (w_{11}, \ldots\ldots, w_{p1})$ defining the first latent component may be used to order the p genes in terms of their relevance for the classification problem.[5] In fact, if the columns of the matrix of predictors X were scaled to unit variance, the F_j-statistic (F-test used in analysis of variance) is a monotonic transformation of the squared weight coefficient w_{j1}^2 ($j = 1, 2, \ldots, p$).[5] A gene selection approach based on several PLS latent components was applied by[2] and.[4] Similarly to this work, in both cases a cross validation was used to choose the number of PLS components. Cross validation technique is useful when the goal is to optimize the predictive power of the model but not specifically in the case of variable selection. It would be interesting to explore the ability of the proposed method to select relevant variables as a function of the number of retained PLS components. A preliminary analysis performed on the Colon dataset revealed that the optimum for variable selection often required a lower number of PLS components than estimated by cross validation (data not shown). Further work is needed to better investigate the relationship between variable selection performance and the number of retained PLS components.

Comparison with other variable selection methods for the two-response case showed that multivariate PLS regression coefficients outperformed Lasso regression coefficients, while obtaining identical performances with PCR regression coefficients. The VIP scores method slightly outperformed all other methods, although it relied on an independent model for each response. In fact, based on the its definition, a VIP score derived by a multivariate PLS regression would not allow to separate the contribution of each predictor to different responses.

In conclusion, this paper gives evidence on the applicability of multivariate PLS regression coefficients in variable selection applied to omics-type of data. This approach is valuable to depict variables that are important to a specific response, while exploiting a comprehensive and compact model as offered by a multivariate PLS. The current study defined also some limits of applicability of the investigated method, as a strong correlation between relevant predictors was an important prerequisite to obtain good performances.

Acknowledgments

We would like to thanks an anonymous referee for his valuable comments that have led to substantial improvement in the paper.

References

1. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of Chemometrics. *Chemom Intell Lab Syst*. 2001;58:109–130.
2. Musumarra G, Barresi V, Condorelli DF, et al. A bioinformatics approach to the identification of candidate genes for the development of new cancer diagnostics. *Biol Chem.* 2003;384:321–327.
3. Wold S. PLS for multivariate linear modeling. In: van de Waterbeemd H, editor. *Chemometric Methods in Molecular Design*. Vol. 2. Weinheim: Verlag Chemie; 1995.
4. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005; 78:102–112.
5. Boulesteix AL. PLS dimension reduction for classification with microarray data. *Stat Appl Genet Mol Biol*. 2004;3(1):33.
6. Wold H. Perspectives in Probability and Statistics. In Gani J (ed). Soft modeling by latent variables: the nonlinear iterative partial least squares approach. London, UK: Academic Press; 1975. p. 520–540.
7. Wold S, Johansson E, Cocchi M. PLS - partial least squares projections to latent structures. In: Kubinyi H, editor. *3D-QSAR in Drug Design, Theory, Methods, and Applications*. Leiden: ESCOM Science Publishers; 1993. p. 523–550.
8. Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Stat*. 2004;32(2):407–499.
9. Hastie T, Efron B. lars: Least Angle Regression, Lasso and Forward Stagewise. R package version 0.9–7. 2007. Accessed on Jan 10, 2009. Available from: http://www-stat.stanford.edu/~hastie/Papers/#LARS.
10. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39:561–577.
11. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology*. 2003;229:3–8.
12. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–537.
13. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Soc*. 2002;97:77–87.
14. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 1999;96(12):6745–6750.
15. Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*. 2001;7:673–679.
16. Selle H, Lamerz J, Buerger K, et al. Identification of novel biomarker candidates by differential peptidomics analysis of cerebrospinal fluid in Alzheimer's Disease. *Comb Chem High Throughput Screen*. 2005;8:801–806.
17. R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria: R Development; 2006. Available from: http://www.R-project.org.
18. Boulesteix AL, Lambert-Lacroix S, Peyre J, et al. Plsgenomics: PLS analyses for genomics. R package version 1.2–2. 2007. Accessed on January 10, 2009. Available from: http://cran.r-project.org/src/contrib/Descriptions/plsgenomics.html.
19. Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998; 95:14863–14868.

Supplementary material

Algorithm to generate the matrix of predictors *X* from a real dataset

Using the *R*′s package *Boost*,[1] an arbitrary number of i.i.d. gene expression profiles, that follow the covariance properties of a dataset of choice, could be generated.

Briefly, the algorithm to generate the *X* matrix work as follows:

1. using a real gene expression dataset of choice it estimates the $(p \times p)$-covariance matrix Σ, as well as the p-dimensional mean vectors $\mu = (\mu_1, \ldots, \mu_p)$
2. Then, for an arbitrary sample size n of choice it repeats independently:
   i. Generate a random vector by the p-dimensional multivariate standard normal distribution

$$z \approx N(0,1_{pxp})$$

   ii. Transform z into a gene expression profile via

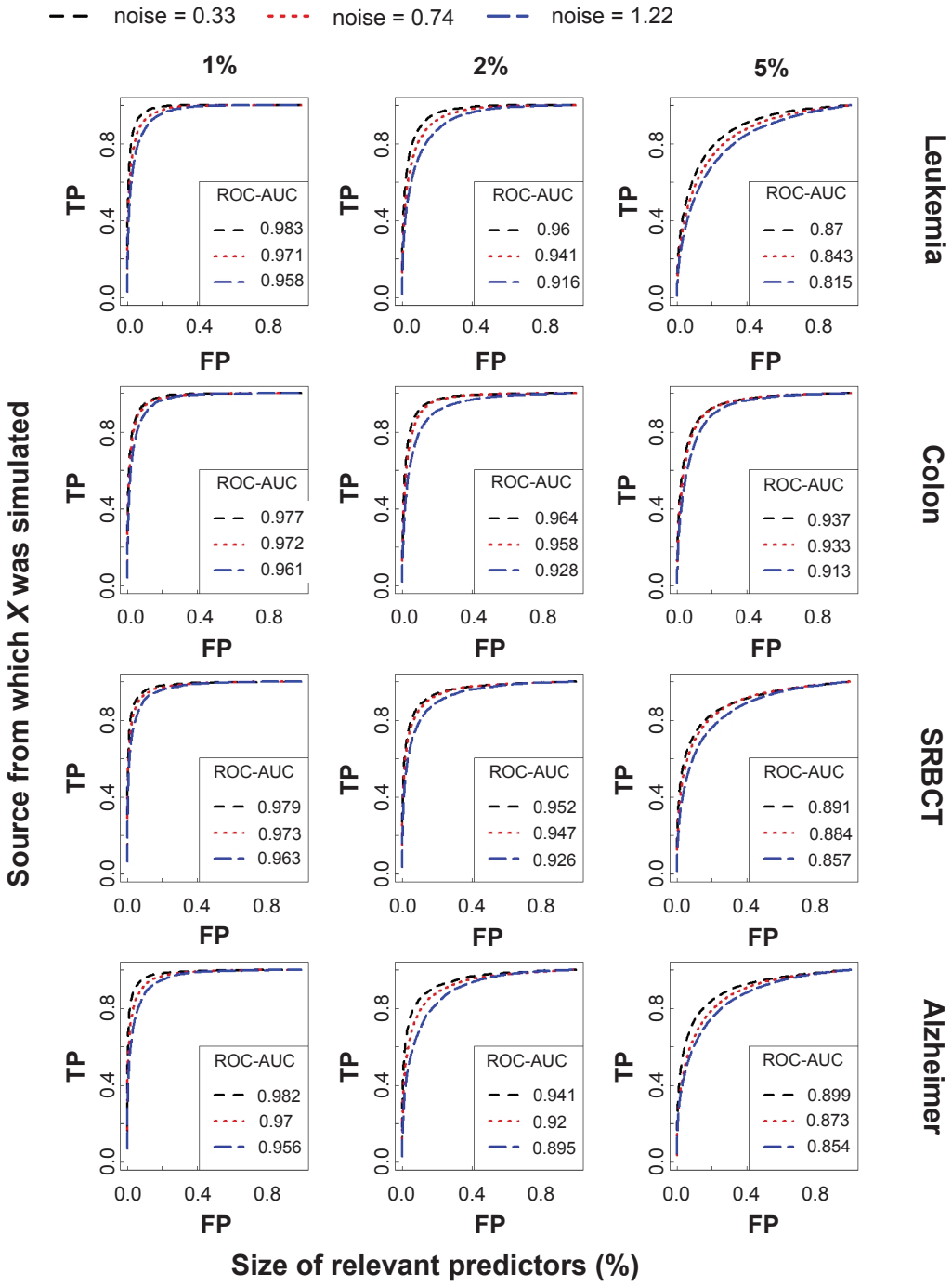
$$x \approx Cz + \hat{\mu}$$

where C is a square root of the covariance matrix Σ, determined by the singular value decomposition.

The above algorithm could be used as well to simulate the covariance structure of LC-MS data.

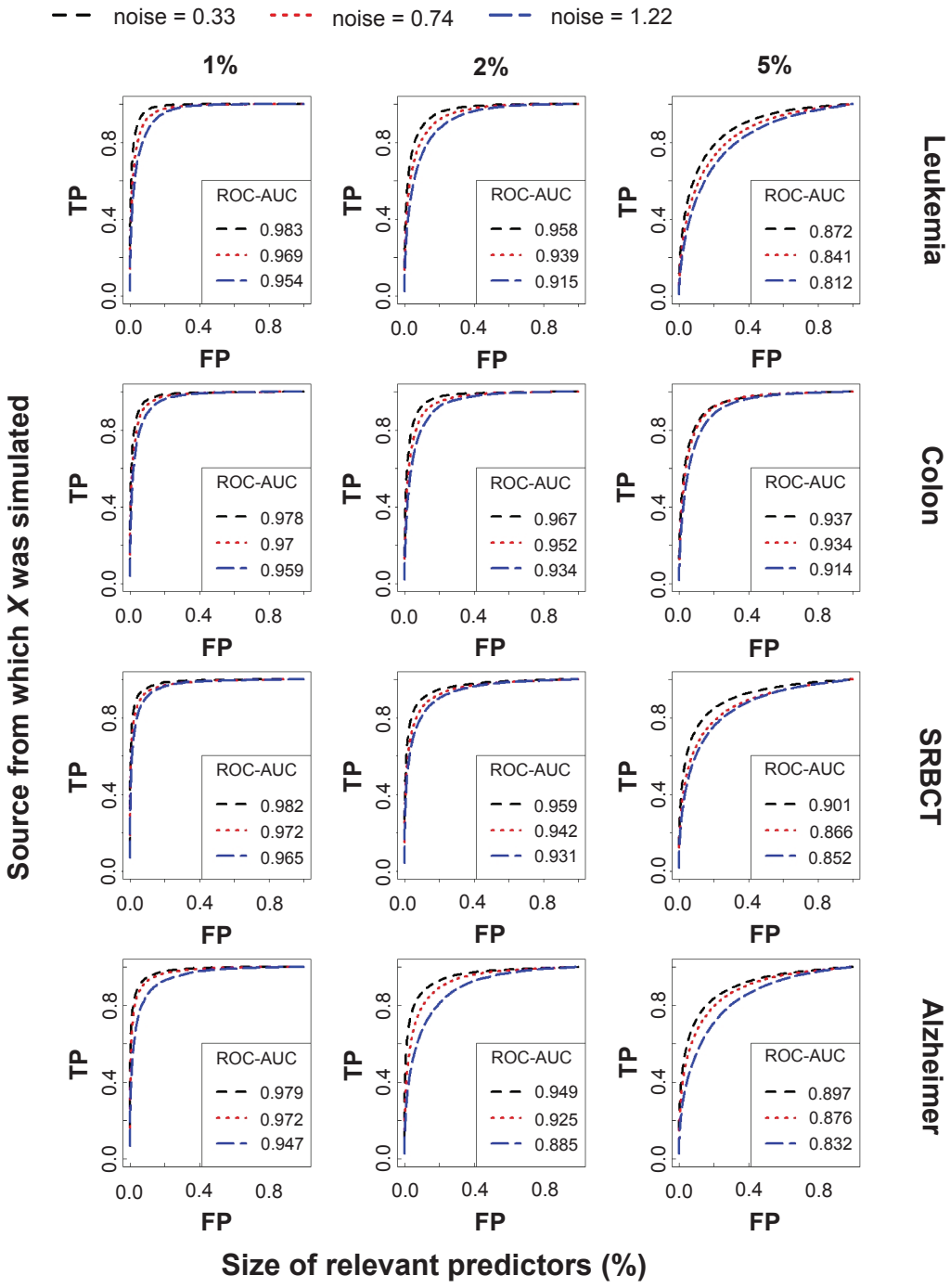
Reference

1. Dettling M. *Bioinformatics*. 2004;20:3583–3593.



Supplementary Figure 1 The performance of PLS regression coefficients, in selecting variables independently for each response of a three-response PLS, is assessed by means of a ROC analysis. ROC curves are evaluated for each of 36 cases of the experimental design. Each curve is an average on the three responses (the average was calculated on 300 ROC curves: 100 replications for each response).

Abbreviations: AUC, area under the curve; FP, false positive; PLS, partial least square; ROC, receiver operating characteristic; SRBCT, small round blue cells tumor; TP, .



Supplementary Figure 2 The performance of PLS regression coefficients, in selecting variables independently for each response of a four-response PLS, is assessed by means of a ROC analysis. ROC curves are evaluated for each of 36 cases of the experimental design. Each curve is an average on the four responses (the average was calculated on 400 ROC curves: 100 replications for each response).

Abbreviations: AUC, area under the curve; FP, false positive; PLS, partial least square; ROC, receiver operating characteristic; SRBCT, small round blue cells tumor; TP, .