BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 1. 32–44
Advance Access publication May 26, 2006
doi:10.1093/bib/bbl016

Partial least squares: a versatile tool for the analysis of high-dimensional genomic data

Anne-Laure Boulesteix and Korbinian Strimmer

Abstract

Partial least squares (PLS) is an efficient statistical regression technique that is highly suited for the analysis of genomic and proteomic data. In this article, we review both the theory underlying PLS as well as a host of bioinformatics applications of PLS. In particular, we provide a systematic comparison of the PLS approaches currently employed, and discuss analysis problems as diverse as, e.g. tumor classification from transcriptome data, identification of relevant genes, survival analysis and modeling of gene networks and transcription factor activities.

Keywords: *partial least squares (PLS); high-dimensional genomic data; gene expression; classification; dimension reduction*

INTRODUCTION

In the last few years, multivariate statistical methods for the analysis of high-dimensional genomic data have been the subject of numerous publications in statistics, machine learning, bioinformatics and biology. A challenging problem connected with these data is that they contain typically many more variables (p, genes and features) than observations (n, gene chips and time points). For instance, it is not uncommon to collect expression data for 20 000 genes using only 10–20 microarrays. Since many traditional multivariate methods are not applicable in this case, predicting, e.g. the survival time or the tumor class of a patient with such high-dimensional data is a difficult and challenging task that requires special techniques such as variable selection or dimension reduction.

In this article, we survey the application of partial least squares (PLS), a powerful yet comparatively unknown approach for analyzing high-dimensional data, to problems in bioinformatics and genomics. The PLS method was first developed by Herman Wold in the 1960s and 1970s to address problems in econometric path modeling, and was subsequently adopted by his son Svante Wold (and many others) in the 1980s for regression problems in chemometric and spectrometric modeling. Early references on path modeling are, e.g. Wold [1–3]. One of the first applications of PLS to regression is Wold *et al.* [4]. Two recent studies [5, 6] describe these early developments and provide a detailed chronological overview. PLS is still a highly active research area from a theoretical point of view; see for instance [7] for recent developments on the connections of PLS with Krylov subspaces and conjugate gradients. PLS started to attract the attention of statisticians only about 15 years ago—see e.g. [8–11]. This was mainly due to the ability of PLS to work very well for data with very small sample sizes and a large number of parameters. Thus, it is only natural that in the last few years this methodology is being successfully applied to problems in genomics and proteomics.

PLS methods are in general characterized by high computational and statistical efficiency. They also offer great flexibility and versatility in terms of the analysis problems that may be addressed.

Corresponding author. Anne-Laure Boulesteix, Department of Medical Statistics and Epidemiology, Technical University of Munich, Ismaningerstrasse 22, D-81675 Munich, Germany. Tel: +49 89 4140-4347; Fax: +49 89 4140-4840; E-mail: anne-laure.boulesteix@tum.de

Anne-Laure Boulesteix is a post-doctoral researcher and consultant in biostatistics at the Technical University of Munich. She received her PhD in statistics in 2005 from the University of Munich, and is generally interested in computational statistics and high-dimensional multivariate data analysis.

Korbinian Strimmer is heading the 'Information Theory and Bioinformatics' group at the Department of Statistics of the University of Munich. His research focuses on statistical learning procedures, complex networks and statistical genomics.

However, the literature of PLS is very diverse because of the existence of a large number of algorithmic variants of PLS, which render it very difficult to understand the principles underlying PLS. It is the aim of this article to fill this gap by, firstly, providing a systematic overview of the available PLS methods and, secondly, reviewing the broad range of their applications to genome data.

The remainder of the article is structured as follows. In 'Methodological Foundations of Partial Least Squares' section, we summarize the main methodological aspects of PLS regression. In 'Applications of Partial Least Squares to High-dimensional Genomic Data' section, various applications of PLS regression to microarray studies are reviewed. 'Outlook and Generalizations of PLS' section is devoted to PLS-based methods that are especially designed for particular types of response variables (for instance, survival time or categorical outcome) and to their practical use in microarray data analysis. A recapitulation of the notations and abbreviations that are used throughout the manuscript can be found in the appendix.

METHODOLOGICAL FOUNDATIONS OF PARTIAL LEAST SQUARES

In this section, we provide an introduction into the mathematics of PLS. In a nutshell, PLS is a dimension reduction approach that is coupled with a regression model. Unlike in similar approaches such as principal component regression, the latent components obtained by PLS are chosen with the response variable of the regression kept in mind.

PLS regression

Suppose we want to predict q continuous response variables $Y_1, \ldots, Y_q$ using p continuous predictor variables $X_1, \ldots, X_p$. The available data sample consisting of n observations is denoted as $(\boldsymbol{x}_i', \boldsymbol{y}_i')_{i=1,\ldots,n}$, where x_i' and y_i' denote the ith observation of the predictor and response variables, respectively. The prime denotes uncentered basic data, as in [9]. Their removal indicates the subtraction of the sample average, i.e.

$$\boldsymbol{x}_i = \boldsymbol{x}_i' - \frac{1}{n}\sum_{s=1}^{n} \boldsymbol{x}_s'$$

$$\boldsymbol{y}_i = \boldsymbol{y}_i' - \frac{1}{n}\sum_{s=1}^{n} \boldsymbol{y}_s'$$

The $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are collected in the $n \times p$ matrix $\mathbf{X}$. Similarly, $\mathbf{Y}$ is the $n \times q$ matrix containing the $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})^{\mathrm{T}}$:

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \ldots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} \boldsymbol{y}_1^{\mathrm{T}} \\ \ldots \\ \boldsymbol{y}_n^{\mathrm{T}} \end{pmatrix}.$$

When $n < p$, the usual regression tools such as classical linear regression, which is often denoted as ordinary least squares (OLS), cannot be applied since the $p \times p$ covariance matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (which can have a maximum rank $n-1$) is singular. In contrast, PLS may be applied also to cases in which $n < p$. PLS regression is based on the basic latent component decomposition:

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathbf{T}} + \mathbf{F}, \tag{1}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \tag{2}$$

where $\mathbf{T}$ is a $n \times c$ matrix giving the latent components for the n observations, $\mathbf{P}$ (of size $p \times c$) and $\mathbf{Q}$ (of size $q \times c$) are matrices of coefficients and $\mathbf{E}$ (of size $n \times p$) and $\mathbf{F}$ (of size $n \times q$) are matrices of random errors. Note that if the given matrices $\mathbf{T}$, $\mathbf{P}$ and $\mathbf{Q}$ satisfy Equations (1) and (2), then so do $\mathbf{T}^\star = \mathbf{T}\mathbf{M}$, $\mathbf{P}^\star = \mathbf{P}(\mathbf{M}^{-1})^{\mathrm{T}}$ and $\mathbf{Q}^\star = \mathbf{Q}(\mathbf{M}^{-1})^{\mathrm{T}}$ for any non-singular $c \times c$ matrix $\mathbf{M}$. Thus, the space spanned by the columns of $\mathbf{T}$ is more important than the columns of $\mathbf{T}$ themselves.

PLS as well as principal component regression and reduced rank regression can all be seen as methods to construct a matrix of latent components $\mathbf{T}$ as a linear transformation of $\mathbf{X}$:

$$\mathbf{T} = \mathbf{X}\mathbf{W}, \tag{3}$$

where $\mathbf{W}$ is a $p \times c$ matrix of weights. In the remainder of the article, the columns of $\mathbf{W}$ and $\mathbf{T}$ are denoted as $\boldsymbol{w}_i = (w_{1i}, \ldots, w_{pi})^{\mathrm{T}}$ and $\boldsymbol{t}_i = (t_{1i}, \ldots, t_{ni})^{\mathrm{T}}$, respectively, for $i = 1, \ldots, c$. For a fixed matrix $\mathbf{W}$, the random variables obtained by forming the corresponding linear transformations of $X_1, \ldots, X_p$ are denoted as $T_1, \ldots, T_c$:

$$\begin{aligned} T_1 &= w_{11}X_1 + \ldots + w_{p1}X_p, \\ \ldots &= \qquad\qquad \ldots \\ T_c &= w_{1c}X_1 + \ldots + w_{pc}X_p. \end{aligned}$$

The latent components are then used for prediction in place of the original variables: once $\mathbf{T}$ is

constructed, $\mathbf{Q}^{\mathrm{T}}$ is obtained as the least squares solution of Equation (1):

$$\mathbf{Q}^{\mathrm{T}} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y}.$$

Finally, the matrix $\mathbf{B}$ of regression coefficients for the model $\mathbf{Y} = \mathbf{XB} + \mathbf{F}$ is given as

$$\mathbf{B} = \mathbf{W}\mathbf{Q}^{\mathrm{T}} = \mathbf{W}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y},$$

and the fitted response matrix $\hat{\mathbf{Y}}$ may be written as

$$\hat{\mathbf{Y}} = \mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y}.$$

If we have a new (uncentered) raw observation $\mathbf{x}_0'$, the prediction $\hat{\mathbf{y}}_0'$ of the response is given by

$$\hat{\mathbf{y}}_0' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i' + \mathbf{B}^{\mathrm{T}}(\boldsymbol{x}_0 - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i').$$

In PLS, dimension reduction and regression are performed simultaneously, i.e. PLS outputs the matrix of regression coefficients $\mathbf{B}$ as well as the matrices $\mathbf{W}$, $\mathbf{T}$, $\mathbf{P}$ and $\mathbf{Q}$, and hence the term PLS regression. In the PLS literature, the columns of $\mathbf{T}$ are often denoted as 'latent variables' or 'scores'. In this study, we prefer the term 'latent components', since in PLS the columns of $\mathbf{T}$ are rather the result of a matrix decomposition than observations of underlying random variables. $\mathbf{P}$ and $\mathbf{Q}$ are often denoted as 'X-loadings' and 'Y-loadings', respectively.

The basic idea of the PLS method is that the response $\mathbf{Y}$ should be taken into account for the construction of the components $\mathbf{T}$. More precisely, the components are defined such that they have high covariance with the response, as outlined in 'Univariate response' and 'Multivariate response' sections. That is why PLS is called a supervized method in contrast to, e.g. principal component analysis (PCA), which does not use the response for the construction of the new components. This feature explains why PLS usually performs better than PCA in prediction problems.

The characterization of the various PLS regression approaches might be done at four different levels:

- the objective function maximized by the $\mathbf{W}$ matrix,
- the $\mathbf{W}$ matrix itself,
- the obtained matrix of regression coefficients $\mathbf{B}$ and
- the algorithm used to compute $\mathbf{W}$.

These four different levels are connected as follows:

- The same $\mathbf{W}$ matrix can maximize several objective functions. But a given objective function is generally satisfied by only one $\mathbf{W}$ matrix (and its opposite–$\mathbf{W}$).
- There might be several algorithms that output the same $\mathbf{W}$ matrix.
- A given $\mathbf{W}$ matrix leads to only one possible matrix of regression coefficients. But two different matrices $\mathbf{W}$ and $\mathbf{W}^{\star}$ can lead to the same regression coefficients if there exists an invertible $c \times c$ matrix $\mathbf{M}$ such that $\mathbf{W}^{\star} = \mathbf{WM}$. Note that, although $\mathbf{W}$ and $\mathbf{W}^{\star}$ lead to the same prediction, they do not necessarily satisfy the same objective function.

Univariate response

In this section, the case of univariate response variables ($q = 1$) is considered. Thus, $\mathbf{Y}$ is a $n \times 1$ matrix, i.e. a vector of length n. Y_1 is denoted as Y in the present section. For a fixed-weight vector $\boldsymbol{w}_{\mathrm{i}} = (w_{1i}, \ldots, w_{pi})^{\mathrm{T}}$, the sample covariance between the response variable Y and the random variable $T_i = w_{1i}X_1 + \ldots + w_{pi}X_p$ can be computed as

$$\widehat{\mathrm{COV}}(Y, T_i) = \frac{1}{n}\boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y},$$

since the matrices $\mathbf{X}$ and $\mathbf{Y}$ contain the centered data. Similarly, for the sample variance of the random variable T_{i}, we have

$$\widehat{\mathrm{VAR}}\,(\mathrm{T}_i) = \boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_i = \frac{1}{n}\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_i$$

and for the sample covariance between T_{i} and T_{j} ($i \neq j$, $i, j = 1, \ldots, c$),

$$\widehat{\mathrm{COV}}(T_i, T_j) = \frac{1}{n}\boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_j = \frac{1}{n}\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_j.$$

In PLS univariate regression, there is only one commonly adopted objective function. The columns $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{\mathrm{c}}$ of the $p \times c$ weight matrix $\mathbf{W}$ are defined such that the squared sample covariance between $\mathbf{Y}$ and the latent components is maximal under the condition that the latent components are mutually empirically uncorrelated. Moreover, the vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{\mathrm{c}}$ are constrained to be of unit length.

Objective function 1: Univariate PLS (PLS1)

For $i = 1, \ldots, c$,

$$\boldsymbol{w}_i = \mathrm{argmax}_{\boldsymbol{w}}\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w},$$

subject to $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i = 1$ and $\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_j = \boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_j = 0$, for $j = 1, \ldots, i-1$,where c is the number of latent components fixed by the user. The maximal number of such latent components that have non-zero covariance with Y is $c_{\max} = \min(n-1, p)$. The weight vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c$ can be computed sequentially via a simple and fast non-iterative algorithm given, e.g. in [12] and denoted as 'algorithm with orthogonal scores' because the matrix $\mathbf{T}^{\mathrm{T}}\mathbf{T}$ is diagonal. Martens and Naes [12] also give another algorithm denoted as 'algorithm with orthogonal loadings', which outputs a different $\mathbf{W}$ matrix. Using this algorithm, one obtains orthogonal loadings instead of orthogonal latent components ($\mathbf{P}^{\mathrm{T}}\mathbf{P}$ is diagonal but not $\mathbf{T}^{\mathrm{T}}\mathbf{T}$). It can be shown [8] that the resulting regression coefficients in matrix $\mathbf{B}$ are the same with both algorithms. Since the orthogonal latent components are easier to interpret than orthogonal loadings, the first algorithm is almost always preferred in the literature. Some statistical aspects of PLS1 regression are discussed by, e.g. [9–11]. From a practical point of view, the objective function of PLS1 can be interpreted as follows. From Equation (4), it is clear that the components constructed in PLS1 have maximal covariance with the response and thus have high predictive power. Moreover, they are not redundant since mutually uncorrelated. The case of multivariate response ($q > 1$) is presented in the following section.

Multivariate response

The case of a multivariate response is more difficult to handle since one has to find latent components which explain all the responses $Y_1, \ldots, Y_q$ simultaneously. There are two main variants for multivariate PLS regression. The first variant is usually denoted as PLS2 in contrast to the univariate method PLS1, or simply PLS. To avoid misunderstandings, we use the term PLS2. The $\mathbf{W}$ matrix corresponding to PLS2 may be obtained via several algorithms. The most well-known are the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm and the Kernel-PLS algorithm, which are implemented in the `R` packages `pls` and `pls.pcr`. Recently, ter Braak and de Jong [13] discovered that the PLS2 maximizes the same expression as Statistically Inspired Modification of PLS (SIMPLS) but with different and less intuitive constraints.

Objective function 2: PLS2

For $i = 1, \ldots, c$,

$$\boldsymbol{w}_i = \mathrm{argmax}_{\boldsymbol{w}} \boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w},$$

subject to $\boldsymbol{w}_i^{\mathrm{T}}(\mathbf{I}_p - \mathbf{W}\mathbf{W}^{+})\boldsymbol{w}_i = 1$ and $\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_j = \boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_j = 0$, for $j = 1, \ldots, i-1$, where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix and $\mathbf{W}^{+}$ is the unique Moore–Penrose inverse of $\mathbf{W}$.

The second important variant of multivariate regression is SIMPLS, which was first introduced by de Jong [14]. In contrast to PLS2, SIMPLS was first developed as an optimality problem. Algorithms were then developed to solve this optimality problem.

Objective function 3: SIMPLS

For $i = 1, \ldots, c$,

$$\boldsymbol{w}_i = \mathrm{argmax}_{\boldsymbol{w}} \boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w},$$

subject to $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i = 1$ and $\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_j = \boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}_j = 0$, for $j = 1, \ldots, i-1$,

The term $\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}$ which is maximized by both PLS2 and SIMPLS is the same as in the univariate case. In the case of a multivariate response ($q > 1$), it can be reformulated as the sum of the squared empirical covariances between $\mathbf{T}$ and $Y_1, \ldots, Y_q$

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w} &= \left((\mathbf{X}\boldsymbol{w})^{\mathrm{T}}\mathbf{Y}\right)^{\mathrm{T}}\left((\mathbf{X}\boldsymbol{w})^{\mathrm{T}}\mathbf{Y}\right) \\ &= n^2 \cdot \sum_{j=1}^{q} \widehat{\mathrm{COV}}\left(T, Y_j\right)^2,\end{aligned}$$

where T is the random variable corresponding to the latent component $\boldsymbol{t} = \mathbf{X}\boldsymbol{w}$. Note that SIMPLS can be seen as a generalization to multivariate response variables of univariate PLS because it has the same criterion $\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}$ and the same constraints. Another equivalent objective function for SIMPLS is often found in the literature, which involves weight vectors for both the response variables and the predictor variables. Based on this formulation, it becomes clear that PLS is connected to classical canonical correlation analysis (CCA). The main difference between the two approaches is that PLS does not maximize correlations but covariances. Thus, PLS does not require the inversion of a $p \times p$ covariance matrix, in contrast to CCA. This feature makes it appropriate for the analysis of high-dimensional data. It can be shown using results from linear algebra [15] that the objective functions 3 and 4 are equivalent.

Objective function 4: SIMPLS (equivalent formulation)

For $i = 1, \ldots, c$

$$(\boldsymbol{w}_i, \boldsymbol{u}_i) = \mathrm{argmax}_{\boldsymbol{w}, \boldsymbol{u}} \boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\boldsymbol{u},$$

subject to $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i = \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_i = 1$ and $\boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_j = \boldsymbol{w}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}_j = 0$, for $j = 1, \ldots, i-1$.

As for PLS2, there exist several algorithms that solve the optimality problem of SIMPLS. One of them is implemented in the function `simpls` from the `R` package `pls.pcr`. A particularity of the `R` function `simpls` is that it returns unit length scores instead of unit length weights (as one would expect when considering objective function 3). By transforming the weights to have unit length, one obtains weights satisfying objective function 3. A user-friendly version of SIMPLS implementing this transformation can be found in the `R` package `plsgenomics` [16].

APPLICATIONS OF PARTIAL LEAST SQUARES TO HIGH-DIMENSIONAL GENOMIC DATA

Regression problems

Any genomic analysis that incorporates a regression model may profit from the application of PLS. Some important recent examples are briefly reviewed in this section.

- A straightforward application of univariate PLS regression to expression data from yeast *Saccharomyces cerevisiae* can be found in [17]. In this study some handpicked gene expression levels are regressed against expression levels of other genes using PLS1 with different numbers of latent components. The magnitude of the obtained regression coefficients are interpreted in terms of interaction strength between genes.
- PLS regression has also been successfully applied to missing values imputation in microarray data by Bras and Menezes [18]. In this approach, the missing values are imputed by PLS regression using all the genes with observed values as predictors. Another reference on PLS imputation in the context of microarray data is Nguyen *et al.* [19].
- Huang *et al.* [20] use PLS regression for a prediction purpose. The aim is to model a continuous variable (LVAD support time) using p gene expression levels as predictors. LVAD stands for 'left mechanical ventricular assist device' and is a successful substitution therapy for heart failure patients waiting for transplantation. Although PLS regression can handle a very large number of predictors and can thus be applied to this problem without adaptation, Huang *et al.* [20] suggest a penalized version of PLS regression (PPLS), which eliminates genes with poor prediction power. Their method is based on the shrinkage of the p regression coefficients obtained by PLS regression. After the shrinkage procedure, a number of genes (depending on the shrinkage parameter Δ) do not contribute anymore to the model. Huang *et al.* [20] suggest to use cross-validation for the selection of both the shrinkage parameter Δ and the number c of latent components used to produce the regression coefficients.
- PLS regression is used by Johansson *et al.* [21] to identify periodically expressed genes. Johansson *et al.* [21] construct a virtual response $\mathbf{Y}$ that represent cyclic behavior with the same periodicity as the cell cycle. The genes that contribute significantly to the PLS regression model are then interpreted as cell-cycle regulated.
- Applications of PLS multivariate regression to other types of data include the prediction of transcription factor activities from combined analysis of gene expression data and chromatin immunoprecipitation (ChIP) data as proposed by Boulesteix and Strimmer [16]. The transcription of genes is regulated by DNA binding proteins, which are known as transcription factors. An issue of interest for biologists is the estimation of the activity levels of these transcription factors. Available data material include microarray data for the potential target genes under different experimental conditions, and 'connectivity' data (e.g. ChIP data) giving the amount of interaction between the transcription factors and the considered genes. Boulesteix and Strimmer [16] assume as the relationship between microarray data and connectivity data the linear structure $\mathbf{Y} = \mathbf{A} + \mathbf{X}\mathbf{B} + \mathbf{F}$, where $\mathbf{Y}$ is the $n \times q$ constant matrix containing the expression levels of n genes (rows) in q conditions (columns), $\mathbf{X}$ is the n x p matrix containing the connectivity information for n genes (rows) and p transcription factors (columns), $\mathbf{A}$ is a $n \times q$ matrix corresponding to the intercepts and $\mathbf{E}$ is a $n \times q$ error matrix. The $p \times q$ matrix $\mathbf{B}$ corresponds to the activity levels of the p transcription factors in the q considered conditions. Thus, the estimation of the transcription factor activities can be formulated as a simple regression problem that is solved in [16] by employing the SIMPLS method. Using PLS in this context allows not only to extract information

on the transcription factors activities but also to identify coherent 'meta-factors' corresponding to the different latent components.
- Other applications of PLS to regression problems in genomic data analysis include, e.g. the prediction of the protein structure (e.g. the helix or strand content using high-dimensional sequence data [22]).

Classification problems

The example above considered only the case of continuous response variables Y. In many studies, however, the response to be predicted is categorical. In other words, Y may take only one of K possible unordered values $Y = 0, \ldots, K-1$. For instance, Y could be the tumor type of a particular cancer patient. If Y is multicategorical ($K > 2$), it has to be transformed before PLS dimension reduction. A simple transformation method consists to convert Y into $K-1$ random variables $Y_1, \ldots, Y_{K-1}$ defined as follows:

$$Y_j = 1 \quad \text{if} \quad Y = j,$$
$$= 0 \text{ otherwise.}$$

Using this transformation, it can be shown that multivariate PLS dimension reduction (almost) leads to the same components as PCA performed on the between-group sample covariance matrix. A collection of properties on this topic as well as mathematical proofs are given in [23]. These properties can be seen as a justification of PLS dimension reduction with categorical variables. Recently, many researchers have considered the PLS methods for classification:

- In two independent comparative studies by Man *et al.* [24] and Huang *et al.* [25], classification based on PLS regression is reported to lead to high prediction accuracy.
- PLS classification analysis for binary response has been investigated by Huang and Pan [26] for leukemia [27] and colon cancer data [28]. Each observation is assigned to one of the two classes 0 or 1, depending on the continuous prediction. Huang and Pan [26] suggest to determine the best number of latent components by leave-one-out cross-validation.
- A similar approach is used in a more applied study by Perez-Enciso and Tenenhaus [29]: various binary outcomes such as (i) before versus after chemotherapy treatment in a case-control study, (ii) estrogen receptor positive versus negative tumors and (iii) tumor type are predicted via PLS discriminant analysis.
- PLS regression is also employed for multiclass classification in [30] for the molecular diagnostic of cancer. Using the software SIMCA, they performed classification with the National Cancer Institute (NCI) data set [31] giving the expression levels of 9605 genes in 60 tumor cell lines of eight different types (leukemia, non-small-cell lung, colon, melanoma, ovarian, breast, central nervous system and renal).
- Other classification studies based on PLS regression can be found in [32–36]. A similar approach based on PLS regression to perform classification in the context of meta-analysis is suggested in [37].

There exists another route to classification using partial least squares, first proposed by Nguyen and Rocke [38, 39] and further studied by Boulesteix [40] and compared with other dimension reduction techniques in [41]. This approach first employs PLS as a dimension reduction method and subsequently uses the PLS latent components as predictors in a classical discrimination method (e.g. logistic regression, linear or quadratic discriminant analysis). To apply this method, one has to choose (i) the number of latent components to be extracted in the dimension reduction step and (ii) the classification method to be used for the classification step.

In Nguyen and Rocke [38, 39], three classification methods are studied: logistic regression, linear discriminant analysis and quadratic discriminant analysis. In [40], the only investigated classification method is linear discriminant analysis. Generally, linear discriminant analysis (LDA) turns out to yield the best classification performance, whereas quadratic discriminant analysis gives worse results. In the extensive comparison study performed by Boulesteix [40], which included many currently employed methods, PLS+LDA turns out to range among the best classification procedures for all the eight studied cancer data sets. According to this study, the most successful other methods are the nearest centroids approach by Tibshirani *et al.* [42] and the support vector machines.

Feature selection

An issue that is tightly connected with the prediction of a clinical outcome is the identification of genes whose expression levels are associated with

the considered outcome. For instance, a physician might want to find out which genes have different expression levels in tumor tissues and normal tissues. The selection of relevant genes is important both for biologists who aim to understand the function of genes and the cell processes and for statisticians who want to apply statistical methods which can handle a restricted number of variables.

In the case of PLS1 dimension reduction (see 'Univariate response' section) applied to binary classification problems (see 'Classification problems' section), the weight vector $\boldsymbol{w}_1 = (w_{11}, \ldots, w_{p1})^{\mathrm{T}}$ defining the first latent component may be used to order the p genes in terms of their relevance for the classification problem [40]. Let F_{j} denote the F-statistic used in analysis of variance and computed from $\mathbf{X}$ for gene j as:

$$F_j = (n-2)\frac{\left(\sum_{k=0}^{1}\sum_{i:\gamma_i=k}\left(\overline{x}_{kj}-\overline{x}_j\right)^2\right)}{\left(\sum_{k=0}^{1}\sum_{i:\gamma_i=k}\left(x_{ij}-\overline{x}_{kj}\right)^2\right)},$$

where

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$$

and

$$\overline{x}_{kj} = \frac{1}{n_k}\sum_{i:\gamma_i=k} x_{ij},$$

with n_k denoting the number of observations from class k in the sample. F_{j} is often used as a selection criterion to order genes in terms of their relevance for the classification problem. Boulesteix [40] proves that F_j is a monotonic transformation of the squared weight coefficient w_{j1}^2 of PLS1 if the columns of the predictor matrix $\mathbf{X}$ have been preliminarily scaled to unit variance. Thus, the ordering of the genes obtained from the weight vector $\boldsymbol{w}_1$ is equivalent to the ordering obtained using the F-statistic, which is one of the most common ordering criteria in microarray data analysis. It shows that PLS dimension reduction and variable selection are in fact two tightly related procedures and also indicates that PLS methods take more information into account than usual univariate gene selection procedures, since they often involve more than one latent component. Similar results might also be obtained in the framework of regression.

A gene selection approach based on several PLS latent components is applied to gene expression data by Musumarra *et al.* [30, 43]. It is based on all the weight vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{\mathrm{c}}$ and implemented in the software package `SIMCA.` The 'variable influence' $\mathrm{VIN}_{\gamma \mathrm{j}}$ of gene j for the γ-th PLS component is defined as a function of $\mathrm{w}_{j\gamma}^2$ and the proportion of the sum of squares explained by the γ-th latent component. Finally, the genes are ordered according to their 'variable importance in the projection' $\mathrm{VIP}_{\mathrm{j}}$, which is defined for each gene j as the sum of the $\mathrm{VIN}_{\gamma \mathrm{j}}$ over the c PLS latent components. An advantage of this approach is that it captures information on the single genes from all the PLS latent components included in the analysis. Thus, it can also discover non-linear patterns which the F-statistic would fail to detect. A major drawback of the VIP index is its lack of theoretical background. One might investigate its connections to the matrix of regression coefficients.

Survival analysis

Another issue of interest in the statistical analysis of gene expression data is the prediction of the survival time Y of diseased patients using their gene expression profiles. In this context, survival data are usually denoted as a triple $(t, \delta, \mathbf{x})$, where:

- t is a continuous variable usually called failure time which equals the time to death Y if $\delta = 1$ or the time to censoring if $\delta = 0$,
- δ is a binary variable, which equals 1 if the death of the patient was observed before censoring and 0 if the patient was still alive at the end of the study,
- $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is a vector of p continuous gene expression levels which are considered as predictor variables.

Standard approaches to predict survival times using continuous predictors such as the proportional hazard regression model (PH model) by Cox [44] may not be applied directly if $n < p$. Various approaches based on the clustering of genes or observations have been proposed, with the inconvenience that the results depend on the chosen clustering algorithm. PLS-based survival analysis is another important family of methods for survival analysis with many predictors.

Nguyen and Rocke [45] suggest a two-stage method that (i) performs univariate PLS with the failure time as response variable and $X_1, \ldots, X_p$ as

predictors and (ii) uses the obtained first latent components as predictors in classical PH regression. They apply their approach to lymphoma data [46] giving the survival time and expression levels of 5622 genes for 40 lymphoma patients and to breast cancer data [47] giving the survival time and expression levels of 3846 genes for 49 breast cancer patients. In this two-step procedure, dimension reduction and prediction using PH regression are performed successively. The specificity of the failure time is not taken into account during the dimension reduction stage: it treats both time to death and time to censoring as the same continuous variable in the dimension reduction step, which is a severe drawback if censoring is non-negligible. Improvements of this approach are proposed in [48–50]. These approaches combine the construction of the successive PLS latent components with PH regression, but in different ways. They are reviewed in 'Outlook and Generalizations of PLS' section which deals with PLS-based methods for special response variables.

Available software

There are currently four R packages that implement partial least squares approaches:

- plsgenomics
  (http://cran.r-project.org/src/contrib/Descriptions/plsgenomics.html)
  This package implements PLS regression (using the function simpls from the pls.pcr package) with user-friendly features such as the choice of the number of components. It also implements the classification method PLS+LDA presented in 'Classification problem' section and discussed by Nguyen and Rocke [38, 39] and Boulesteix [40] as well as the ridge PLS method [51] mentioned in 'PLS and generalized linear models' section.
- pls.pcr
  (http://cran.r-project.org/src/contrib/Descriptions/pls.pcr.html)
  This package implements the two main variants of multivariate PLS regression SIMPLS and PLS2 as well as PCR.
- pls
  (http://cran.r-project.org/src/contrib/Descriptions/pls.html)
  This package is an extension of the earlier package pls.pcr including, e.g. various plot functions and a formula interface.
- gpls
  (http://cran.r-project.org/src/contrib/Descriptions/gpls.html)
  This package implements the classification method using generalized PLS [52] mentioned in 'PLS and generalized linear Models' section.
- plss
  (http://www.math.univ-montp2.fr/~durand/ProgramSources.html)
  These programs implement PLS regression based on splines transformations of the predictors [53]. They work only under R for Windows.

Other software

- Classification with PLS regression (PLS-DA), (DA, discriminant analysis) is implemented in the software tool SIMCA.
  (http://www.umetrics.com/default.asp/pagename/software_simcap/c/3/).
- The SAS procedure PLS implements several dimension reduction methods such as PCR, Reduced Rank Regression (RRR) and PLS. The two main versions of multivariate PLS (SIMPLS and PLS2) are available. For PLS2, one may specify the algorithmic variant as an option, for instance NIPALS.
  (http://support.sas.com/rnd/app/da/new/dapls.html)
- The PLS Toolbox (by Eigenvector Research Incorporated) for use with MATLAB
  (http://software.eigenvector.com/toolbox/3_5/index.html)
  includes a wide range of methods for multivariate statistical analysis, some of which are based on PLS regression. In particular, it includes the function plsda, which performs classification (class prediction) based on SIMPLS or PLS2 regression.
- The software tool Unscrambler
  (http://www.camo.com/rt/Products/Unscrambler/unscrambler.html)
  also implements multivariate PLS1 and multivariate regression (PLS2) and PLS-DA.

OUTLOOK AND GENERALIZATIONS OF PLS

So far, we have considered applications of PLS regression to various biological problems. However, applying a regression method designed for continuous responses to categorical responses or

performing dimension reduction with survival data without taking censoring into account is unappealing, although it is reported to give good results in many cases. In this section, we review methods that use the principle of PLS regression but adapt it to handle special types of responses such as survival time or categorical outcome. These methods can be divided into two categories. In the first category of methods, the structure of the univariate PLS regression algorithm remains unchanged, but the coefficients used to construct the latent components are modified. In the second category of methods, the PLS algorithm is embedded into a complex generalized regression procedure. Both approaches can be applied to, e.g. survival analysis and classification. In the following section, we consider only the univariate case, i.e. $\mathbf{Y}$ is a $n \times 1$ matrix (n vector).

Modification of the latent components in PLS regression

Let us consider objective function 1. Some calculation using the Lagrange multiplier method yields

$$\boldsymbol{t}_1 = \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}/||\mathbf{X}^{\mathrm{T}}\mathbf{Y}||.$$

In the most usual PLS1 algorithm, the weight vectors $\boldsymbol{t}_2, \ldots, \boldsymbol{t}_c$ are built sequentially in a similar way as $\boldsymbol{t}_1$, except that $\mathbf{X}$ and $\mathbf{Y}$ are replaced by deflated matrices. With $\boldsymbol{t}_1^{\mathrm{T}} = (t_{11}, \ldots, t_{n1})$ and x_{ij} denoting the element of $\mathbf{X}$ at row i and column j, simple transformations lead to

$$t_{i1} \propto \sum_{j=1}^{p} \widehat{\mathrm{COV}}\,(Y, X_j)\, x_{ij}$$

$$\propto \sum_{j=1}^{p} \widehat{\mathrm{VAR}}\,(X_j)\, \beta_j\, x_{ij},$$

where β_j is the least squares regression coefficient obtained by regressing Y against X_j. The subsequent vectors $\boldsymbol{t}_2, \ldots, \boldsymbol{t}_c$ may be expressed in a similar way using deflated matrices. Several studies are based on the idea that β_j is not an optimal choice when Y is a binary or survival variable. Li and Gui [50] suggest to replace β_j by the regression coefficient of X_j obtained via Cox regression analysis, thus taking the specificity of the response variable Y into account. For the construction of $\boldsymbol{t}_1$, Y is regressed against X_j. For the construction of $\boldsymbol{t}_j$, $j > 1$, Y is regressed against X_j and the j-1 first latent components. A similar approach is proposed by Bastien [54] and studied from a methodological point of view in [55]. The idea to replace a linear regression coefficient by a Cox regression coefficient also inspired another method denoted as 'MPLS': Nguyen [48] gives a different non-sequential expression of the PLS1 latent components $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_c$ involving eigenvectors of the matrices $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ (see [56] for details). This complex expression also contains a linear regression coefficient, which Nguyen [48] replaces by a Cox regression coefficient. The same approach is also used in the context of binary classification [56] and denoted as 'PLSM2'.

A related approach denoted as PLS logistic regression is used in [57] to map complex trait genes using gene expression data. In this setting, the response is a categorical genetic trait and the latent components $\mathbf{t}_2, \ldots, \mathbf{t}_c$ are constructed based on the regression coefficients estimated from a logistic regression model. Perez-Enciso *et al.* [57] demonstrate the potentialities of this approach based on an extensive simulation study.

PLS and generalized linear models

Marx [58] proposes an extension of the concept of PLS regression into the framework of generalized linear models. This approach, which is denoted as iteratively reweighted partial least squares (IRPLS or IRWPLS), embeds the univariate PLS regression algorithm into the iterative steps of the usual Iteratively Reweighted Least Squares algorithm [59] for generalized linear models, resulting in two nested loops. The loops are iterated a fixed number of times or until a convergence criterion is reached. This apparently appealing approach has a major drawback in practical microarray data analysis: convergence is never reached if $\mathbf{X}$ is full row-rank, which is most often the case in high-dimensional microarray data with $n \ll p$ [51]. The IRPLS method as well as a few adaptations overcoming the convergence problem have been applied both to survival analysis and classification. Binary classification is one of the most common applications of generalized linear models and of Marx's IRPLS algorithm. To our knowledge, the IRPLS algorithm has never been applied directly to classification with microarray data. However, it has inspired at least two recent papers on the generalization of PLS regression to categorical response variables.

The first approach is proposed by Ding and Gentleman [52] and can be seen as an adaptation of Marx's IRPLS method which solves the problem of separation. As already mentioned in 'Classification

problems' section, infinite parameter estimates can occur in binary logistic regression when the two classes are completely or quasi-completely separated [60]. Firth [61] suggests a procedure to remove the first-order term of the asymptotic bias of maximum likelihood estimates in Generalized Linear Models (GLMs). The procedure is based on a modified score function which, when applied to logistic regression, guarantees finite estimates [62]. The binary classification method obtained by using the Firth's modified score function in place of the usual score function in the IRPLS algorithm is denoted as IRWPLSF by Ding and Gentleman [52]. They also propose a generalization of the method to multicategorical response variables, which is based on the multinomial logit model and denoted as MIRWPLSF. The IRWPLSF and MIRWPLSF are reported to achieve a slightly better classification performance than usual classification methods such as nearest neighbors or SVM on the colon cancer data [28] and on the NCI cancer data [31]. The second approach to modify Marx's IRPLS is suggested in [51]: the procedure embeds a PLS step into ridge penalty logistic regression and might also be generalized to multicategorical responses. This method is applied with success to the colon cancer data [28], the leukemia data [27] and the prostate cancer data [63].

Another classical application of generalized linear models and IRPLS is survival analysis. As suggested in [64], Park *et al*. [49] transform the failure time problem into a generalized linear regression problem with logarithmic link function. They propose to use the IRPLS estimation method for generalized linear regression [58]. In contrast to the two-stage scheme developed in [45], this method takes censoring explicitly into account. The choice of the number of components is done via a cross-validation procedure which suggests to use $c = 1$ for the lung cancer data set [65]. According to Park *et al*. [49] convergence is achieved in a few steps. However, this property seems to be controversial and lack of convergence problems are invoked as a drawback of the method in the more recent paper by Li and Gui [50].

CONCLUSIONS

The microarray 'revolution' has lead to an enormous increase in the availability of high-dimensional biomedical data. Classical multivariate methods are not applicable to these 'small n, large p' data sets. In this article we have reviewed the PLS approach to regression and dimension reduction that is perfectly suited for analysing this kind of data.

Specifically, PLS has several advantages over many competing approaches:

- It automatically performs variable selection.
- It can be applied to a diverse set of tasks, including classification, survival analysis and modeling of transcription factors activities.
- It is statistically very efficient.
- Moreover, it is computationally very fast, which renders it practical for application to large data sets.

As outlined in 'Application of Partial Least Squares to High-dimensional Genomic Data' and 'Outlook and Generalizations of PLS' sections of this review, at present most reported applications of the PLS method to genomic data focus on the analysis of microarray data from gene expression experiments. The key advantages that characterize the PLS methodology are versatility and flexibility. On the one hand, it can be directly applied to various types of data of any dimensions for different prediction or imputation problems. On the other hand, PLS algorithms adapt easily to a broad range of questions and thus serve as a flexible basis for the development of novel tools for the analysis biological data. In short, we expect that with the advent of proteomics data, e.g. from mass spectrometric experiments, PLS will in the future also play a major role for analysing many other kinds of high-dimensional omics data.

Key Points

- PLS is an efficient statistical prediction tool that is especially appropriate for small sample data with many (possibly correlated) variables.
- PLS is fast, easy to implement and does not necessitate any preliminary feature selection.
- The problems that may be addressed by the PLS method are very diverse and include, e.g. tumor diagnosis, survival analysis, and modeling of regulation network.

References

1. Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR (ed). *Multivariate Analysis*. New York: Academic Press, 1966; 391–420.
2. Wold H. Nonlinear Iterative Partial Least Squares (NIPALS) modeling: some current developments.

In: Krishnaiah PR (ed). *Multivariate Analysis*. New York: Academic Press, 1973;383–407.

3. Wold H. Path models with latent variables: the NIPALS approach. In: Blalock HM (ed). *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*. New York: Academic Press, 1975.
4. Wold S, Ruhe A, Wold H, *et al*. Collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J Sci Comput Stat* 1984;**5**:735–43.
5. Martens H. Reliable and relevant modeling of real world data: a personal account of the development of PLS regression. *Chemom Intell Lab Syst* 2001;**58**:85–95.
6. Wold S. Personal memories of the early PLS development. *Chemom Intell Lab Syst* 2001;**58**:83–4.
7. Phatak A, Dehoog F. Exploiting the connection between PLSR, Lanczos, and conjugate gradients: alternative proofs of some properties of PLSR. *J Chemom* 2002;**16**:361–7.
8. Helland I. On the structure of Partial Least Squares. *Comm Stat Simul Comp* 1988;**17**:581–607.
9. Stone M, Brook RJ. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *J Roy Stat Soc B* 1990;**52**:237–69.
10. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993;**35**: 109–35.
11. Garthwaite PH. An interpretation of partial least squares. *J Am Stat Assoc* 1994;**89**:122–7.
12. Martens H, Naes T. *Multivariate Calibration*. New York: Wiley, 1989.
13. Braak CJF, de Jong S. The objective function of partial least squares. *J Chemom* 1998;**12**:41–54.
14. Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemom Intell Lab Syst* 1993;**18**:251–63.
15. Rao CR. *Linear Statistical Inference and its Application*. New York: Wiley, 1993.
16. Boulesteix AL, Strimmer K. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor Biol Med Model* 2005;**2**:23.
17. Datta S. Exploring the relationships in gene expressions: a partial least squares approach. *Gene Expression* 2001;**9**: 257–64.
18. Bras LP, Menezes JC. Dealing with gene expression missing data. *IEE Syst Biol* 2006;**153**:105–19.
19. Nguyen DV, Wang N, Caroll RJ. Evaluation of missing value estimation for microarray data. *J Data Sci* 2004;**2**: 347–70.
20. Huang X, Pan W, Park S, *et al*. Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics* 2004;**20**:888–94.
21. Johansson D, Lindgren P, Berglund A. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics* 2003;**19**:467–73.
22. Clementi M, Clementi S, Cruciani G, *et al*. Robust multivariate statistics and the prediction of protein secondary structure content. *Protein Eng* 1997;**10**:747–9.
23. Barker M, Rayens W. Partial least squares for discrimination. *J Chemom* 2003;**17**:166–73.
24. Man MZ, Dyson G, Johnson K, *et al*. Evaluating methods for classifying expression data. *J Biopharm Stat* 2004;**14**:1065–84.
25. Huang X, Pan W, Grindle S, *et al*. A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics* 2005;**6**:205.
26. Huang X, Pan W. Linear regression and two-class classification with gene expression data. *Bioinformatics* 2003;**19**:2072–8.
27. Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**:531–7.
28. Alon U, Barkai DA, Notterman K. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci* 1999;**96**:6745–50.
29. Perez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares approach. *Hum Genet* 2003;**112**:581–92.
30. Musumarra G, Barresi V, Condorelli DF, *et al*. Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis. *J Chemom* 2004; **18**:125–32.
31. Ross DT, Scherf U, Eisen MB, *et al*. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000;**24**:227–34.
32. Alaiya AA, Franzen B, Hagman A, *et al*. Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. *Int J Cancer* 2000;**86**:731–6.
33. Musumarra G, Condorelli DF, Scire S, *et al*. Shortcuts in genome-scale cancer pharmacology research from multivariate analysis of the National Cancer Institute gene expression data base. *Biochem Pharmacol* 2001;**62**:547–53.
34. Cho JH, Lee D, Park JH, *et al*. Optimal approach for classification of acute leukemia subtypes based on gene expression data. *Biotech Progress* 2002;**18**:847–54.
35. Tan Y, Shi L, Tong W, *et al*. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput Biol Chem* 2004;**28**:235–44.
36. Modlich O, Prisack HB, Munnes M, *et al*. Predictors of primary breast cancers responsiveness to preoperative epirubicin//cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. *J Transl Med* 2005;**3**:32.
37. Huang X, Pan W, Han X, *et al*. Borrowing information from relevant microarray studies for sample classification using weighted partial least squares. *Comput Biol Chem* 2005; **29**:204–11.
38. Nguyen DV, Rocke D. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002;**18**:39–50.
39. Nguyen DV, Rocke D. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 2002;**18**:1216–26.
40. Boulesteix AL. PLS dimension reduction for classification with high-dimensional microarray data. *Stat Appl Genet Mol Biol* 2004;**3**:33.

41. Dai JJ, Lieu L, Rocke D. Dimension reduction for classification with gene expression data. *Stat Appl Genet Mol Biol* 2006;**5**:6.

42. Tibshirani R, Hastie T, Narasimhan B, *et al*. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci* 2002;**99**:6567–72.

43. 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**:321–7.

44. Cox DR. Regression models and life-tables (with discussion). *J Roy Stat Soc B* 1972;**34**:187–220.

45. Nguyen DV, Rocke D. Partial least squares proportional hazards regression for application to DNA microarray survival data. *Bioinformatics* 2002;**18**:1625–32.

46. Alizadeh AA, Eisen MB, Davis RE, *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;**403**:503–11.

47. Sorlie T, Perou CM, Tibshirani R, *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci* 2001;**98**: 10869–74.

48. Nguyen DV. Partial least squares dimension reduction for microarray gene expression data with a censored response. *Math Biosci* 2005;**193**:119–37.

49. Park PJ, Tian L, Kohane IS. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* 2002;**20**:208–15.

50. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* 2004;**20**:208–15.

51. Fort G, Lambert-Lacroix S. Classification using partial least squares with penalized logistic regression. *Bioinformatics* 2005;**21**:1104–11.

52. Ding B, Gentleman R. Classification using penalized partial least squares. *J Comput Graph Stat* 2005;**14**:280–98.

53. Durand JF. Local polynomial additive regression through PLS and splines: PLSS. *Chemom Intell Lab Syst* 2001;**58**: 235–46.

54. Bastien P. PLS-Cox model: application to gene expression data. In: Proceedings COMPSTAT'04. Springer: Physica-Verlag, 2004;655–62.

55. Bastien P, Esposito-Vinzi V, Tenenhaus M. PLS generalized linear regression. *Comput Stat Data Anal* 2005;**48**:17–46.

56. Nguyen DV, Rocke D. On partial least squares dimension reduction for microarray-based classification: a simulation study. *Comput Stat Data Anal* 2004;**46**:407–25.

57. Perez-Enciso M, Toro MA, Tenenhaus M, *et al*. Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. *Genetics* 2003;**164**:1597–606.

58. Marx BD. Iteratively reweighted partial least squares. *Technometrics* 1996;**38**:374–81.

59. Green P. Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *J Roy Stat Soc B* 1984;**46**:149–92.

60. Albert A, Anderson J. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984;**71**:1–10.

61. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993;**80**:27–38.

62. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002;**21**:2409–19.

63. Singh D, Febbo PG, Ross K, *et al*. Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell* 2002;**1**:203–9.

64. Whitehead J. Fitting Cox's regression model to survival data using GLIM. *J Roy Stat Soc C* 1980;**29**:268–75.

65. Bhattacharjee A, Richards WG, Staunton J, *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci* 2001;**98**:13790–5.

APPENDIX

List of abbreviations

Term	Signification	Introduced in sections
PLSI	Univariate PLS	Univariate response
PLS2	Multivariate PLS (first)	Multivariate response
SIMPLS	Multivariate PLS (second)	Univariate response
OLS	Ordinary Least Squares	
PCR	Principal Component Regression	
PCA	Principal Component Analysis	
RRR	Reduced Rank Regression	
PLS+LDA	Two-step classification procedure consisting of PLS dimension reduction and LDA	Classification problems
IRPLS	Marx's Iteratively Reweighted PLS	PLS and generalized linear models
$\mathbf{X} = (x_{ij})_{i=1,\ldots,n,\, j=1,\ldots,p}$	$n \times p$ matrix of predictors	PLS regression
$\mathbf{Y} = (y_{ij})_{i=1,\times,n, j=1,\ldots,q}$	$n \times q$ response matrix	PLS regression
$X_1, \ldots, X_p$	Uncentered predictor variables (random variables)	PLS regression
$Y_1, \ldots, Y_q$	Uncentered response variables (random variables)	PLS regression
$(x_i', y_i')_{i=1,\ldots,n}$	Uncentered sample	PLS regression
$(\mathbf{x}_i, \mathbf{y}_i)_{i=1,\ldots,n}$	Centered sample	PLS regression
$\mathbf{w}_j = (w_{1j}, \ldots, w_{pj})^T$	Weight vector defining the j-th latent component	PLS regression
$\mathbf{t}_j = (t_{1j}, \ldots, t_{nj})^T$	j-th latent component	PLS regression
$\mathbf{T} = [\mathbf{t}_1, \ldots, t_c]$	$n \times c$ matrix of latent components	PLS regression
$\mathbf{W} = [w_1, \ldots, w_c]$	$p \times c$ matrix of weights	PLS regression
T_j, $j = 1, \ldots, c$	(Uncentered) random variable corresponding to t_j	PLS regression
P	$p \times c$ matrix of X-loadings	PLS regression
Q	$q \times c$ matrix of Y-loadings	PLS regression
E	$n \times p$ error matrix	PLS regression
F	$n \times q$ error matrix	PLS regression
B	$p \times q$ matrix of regression coefficients	PLS regression