Boulesteix, Strimmer:

Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data

Sonderforschungsbereich 386, Paper 457 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data

Anne-Laure Boulesteix* and Korbinian Strimmer†

17 October 2005

Corresponding author:

Anne-Laure Boulesteix
Tel.: +49 89 2180-3466
Fax: +49 89 2180-5308
Email: anne-laure.boulesteix@stat.uni-muenchen.de

Keywords: Partial least squares (PLS), regression, dimension reduction, microarray, gene expression analysis.

Short title: PLS Approaches for Genomic Data

*Seminar for Applied Stochastics, Department of Statistics, University of Munich, Akademiestrasse 1, D-80799 Munich, Germany.

†Statistical Genetics and Bioinformatics Group, Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany.

Abstract

Partial Least Squares (PLS) is a highly efficient statistical regression technique that is well suited for the analysis of high-dimensional genomic data. In this paper we review the theory and applications of PLS both under methodological and biological points of view. Focusing on microarray expression data we provide a systematic comparison of the PLS approaches currently employed, and discuss problems as different as tumor classification, identification of relevant genes, survival analysis and modeling of gene networks.

1 Introduction

In the last few years, multivariate statistical methods for microarray data analysis have been the subject of numerous publications in statistics, machine learning, bioinformatics and biology. A challenging problem connected with transcriptome data is that they contain typically many more variables (p, genes) than observations (n, gene chips, time points). For instance, it is not uncommon to collect expression data for 20,000 genes using only 10-20 microarrays. Since most traditional multivariate techniques 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.

A powerful yet comparatively unknown approach for analyzing high-dimensional microarray data analysis is *supervised* dimension reduction based on Partial Least Squares (PLS). PLS is also known as a regression method, since the obtained latent components may be used instead of the original variables in regression to overcome the dimensionality problem $n \ll p$. As a supervised approach, it uses the response variable of interest in the dimension reduction step, which often makes it more efficient in prediction problems than the unsupervised Principal Component Analysis (PCA) approach [1]. In contrast to other supervised dimension approaches such as sufficient dimension reduction [2, 3, 4], it is applicable and very fast even if the number of variables is much larger than the number of observations. As an alternative to dimension reduction, most authors cope with the high-dimensionality of microarray data by selecting the variables (genes) of interest preliminarily to their analysis. Using this approach, a large amount of information is systematically excluded from the analysis and interactions and correlations between genes are often omitted. Moreover, the results of the statistical analysis depend largely on the variable selection procedure and on the number of selected variables, which is most often chosen on a purely heuristic basis. Thus, global dimension reduction methods such as PLS are especially appropriate to deal with high-dimensional microarray data.

PLS methods are characterized by (i) a high computational efficiency, (ii) a great flexibility and versatility in terms of the addressed concrete problems, (iii) the existence of a large variety of diverse algorithmic variants. The points (i) and (ii) render PLS methods very attractive for the analysis of microarray data. It is the aim of this paper to address point (iii) by providing a systematic overview over available PLS method and to review the broad range of their applications to genome data.

In this paper, we review applications of PLS methods to the analysis of microarray data both under the methodological and biological points of view. In Section 2, we summarize the main methodological aspects of PLS regression. In Section 3, various applications of PLS regression to microarray studies are reviewed. Section 4 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.

2 PLS dimension reduction and regression

2.1 History and framework

Two recent papers [5, 6] about the early developments of PLS regression give a chronological overview of how PLS regression emerged from Herman Wold's work on multi-block path modeling. The term 'Partial Least Squares' first referred to a general approach developed by Herman Wold in the 60s and 70s, which is based on successive least squares fits and used in the context of path modeling with latent variables. Early references are, e.g., Wold [7], Wold [8], or Wold [9]. The connection between PLS path modeling and the statistical LISREL approach is studied in Schneeweiss [10]. Applications of partial least squares methods to regression problems are first proposed in the early 80s and focus on the analysis of high-dimensional chemometric data. Most methodological papers on PLS regression and dimension reduction can be found in the journals *Journal of Chemometrics* and *Chemometrics and Intelligent Laboratory Systems*. PLS regression is studied from the point of view of statisticians in, e.g., Helland [11], Stone and Brooks [12] and Frank and Friedman [13]. It is often described as a data analysis tool rather than a proper statistical method as it lacks an underlying probabilistic model [14]. Nevertheless, its efficiency when applied to very high-dimensional data is unquestionable. However, the many variants of PLS methods render them very confusing and difficult to understand at first sight. Sections 2.3 and 2.4 give an overview of the most common approaches in the case of univariate and multivariate responses, respectively.

There have been several attempts to place PLS into a global regression or dimension reduction framework. Dimension reduction methods that are related to PLS in terms of the objective function include Principal Component Regression (PCR) and Reduced Rank Regression (RRR). PLS, OLS and PCR and the Ridge Regression (RR) method can also be tied together within a continuum regression framework [12]. The connections of PLS to PCR, OLS, RRR and RR are briefly reviewed in Section 2.5.

2.2 Introduction to 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 $(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{y}}_i)_{i=1,\ldots,n}$, where $\dot{\boldsymbol{x}}_i \in \mathbb{R}^p$ and $\dot{\boldsymbol{y}}_i \in \mathbb{R}^q$ denote the i-th observation of the predictor and response variables, respectively. The dots denote uncentered basic data, as in Stone and Brooks [12]. Their removal indicates the subtraction of the sample average, i.e.

$$\begin{aligned} \boldsymbol{x}_i &= \dot{\boldsymbol{x}}_i - \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \dot{\boldsymbol{x}}_j \\ \boldsymbol{y}_i &= \dot{\boldsymbol{y}}_i - \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \dot{\boldsymbol{y}}_j. \end{aligned}$$

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

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

When $n < p$, the usual regression tools such as classical linear regression, which is often denoted as Ordinary Least Squares (OLS) can not be applied since the $p \times p$ covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$ (which has rank at most $n - 1$) is singular. In contrast, PLS may be applied also to cases where $n < p$. PLS regression is based on the basic latent component decomposition

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Q}^T + \boldsymbol{F}, \tag{2}$$

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

PLS, as well as PCR and RRR can all be seen as methods to construct a matrix of latent components $\boldsymbol{T}$ as a linear transformation of $\boldsymbol{X}$:

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

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

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

The latent components are then used for prediction in place of the original variables: once $\boldsymbol{T}$ is constructed, $\boldsymbol{Q}$ is obtained as the least squares solution of equation (2):

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

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

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

and the fitted response matrix $\widehat{\boldsymbol{Y}}$ may be written as

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y}.$$

If we have a new raw observation $\dot{\boldsymbol{x}}_0$, the prediction $\widehat{\dot{\boldsymbol{y}}_0}$ of the response is given by

$$\widehat{\dot{\boldsymbol{y}}_0} = \frac{1}{n}\sum_{j=1}^{n}\dot{\boldsymbol{y}}_j + \boldsymbol{B}^T(\dot{\boldsymbol{x}}_0 - \frac{1}{n}\sum_{j=1}^{n}\dot{\boldsymbol{x}}_j).$$

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

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

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

These four different levels are connected as follows:

- The same $\boldsymbol{W}$ matrix can maximize several objective functions. For instance, the SIMPLS objective function 3 (SIMPLS) might be reformulated as in objective function 4. But a given objective function is generally satisfied by only one $\boldsymbol{W}$ matrix (and its opposite $-\boldsymbol{W}$).
- There might be several algorithms that output the same $\boldsymbol{W}$ matrix.
- A given $\boldsymbol{W}$ matrix leads to only one possible matrix of regression coefficients. But two different matrices $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ can lead to the same regression coefficients, if there exists an invertible $p \times p$ matrix $\boldsymbol{M}$ such that $\tilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{M}$. Note that, although $\boldsymbol{W}$ and

$\tilde{\boldsymbol{W}}$ lead to the same prediction, they do not necessarily satisfy the same objective function.

2.3 Univariate response

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

$$\widehat{\mathrm{Cov}}(Y, T_i) = \frac{1}{n}\boldsymbol{w}_i^T \boldsymbol{X}^T \boldsymbol{Y}, \tag{4}$$

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

$$\widehat{\mathrm{Var}}(T_i) = \frac{1}{n}\boldsymbol{w}_i^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}_i = \frac{1}{n}\boldsymbol{t}_i^T \boldsymbol{t}_i.$$

and for the sample covariance of T_i and T_j ($i \neq j$, $i, j = 1, \ldots, c$)

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

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

Objective function 1 Univariate PLS (PLS1)
For $i = 1, \ldots, c$

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{X}\boldsymbol{w}$$

subject to $\boldsymbol{w}_i^T \boldsymbol{w}_i = 1$ *and* $\boldsymbol{t}_i^T \boldsymbol{t}_j = \boldsymbol{w}_i^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}_j = 0$, *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 which have non-zero covariance with Y is $c_{max} = \min(n, p)$. 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 Martens and Naes [15] and denoted as "algorithm with orthogonal scores", because the matrix $\boldsymbol{T}^T\boldsymbol{T}$ is diagonal. Martens and Naes [15] also give another algorithm denoted as "algorithm with orthogonal loadings" which outputs a different $\boldsymbol{W}$ matrix. Using this algorithm, one obtains orthogonal loadings instead of

orthogonal latent components ($\boldsymbol{P}^T\boldsymbol{P}$ is diagonal but not $\boldsymbol{T}^T\boldsymbol{T}$). It can be shown [11] that the resulting regression coefficients in matrix $\boldsymbol{B}$ 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 univariate PLS regression are discussed, e.g., in Stone and Brooks [12], Garthwaite [14] and Frank and Friedman [13]. The case of multivariate response ($q > 1$) is presented in the following section.

2.4 Multivariate response

The case of 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 $\boldsymbol{W}$ matrix corresponding to PLS2 may be obtained via several algorithms. The most well-known are the NIPALS algorithm and the Kernel-PLS algorithm which is implemented in the `R` packages `pls` and `pls.pcr`. Recently, ter Braak and de Jong [16] discovered that the PLS2 maximizes the same expression as SIMPLS, but with different -and less intuitive constraints.

Objective function 2 PLS2

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{w})$$

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

The second important variant of multivariate regression is SIMPLS (Statistically Inspired Modification of PLS), which is first introduced by de Jong [17]. 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

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{w})$$

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

The term $\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{w}$ 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 T and $Y_1, \ldots, Y_q$:

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

where T is the random variable corresponding to the latent component $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$. Note that SIMPLS can be seen as a generalization to multivariate response variables of univariate PLS, because it has the same criterion $\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{X}\boldsymbol{w}$ 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. Since this objective function is the most common one, we give it here for exhaustivity, although objective function 3 is certainly more easy to interpret. It can be shown using results from linear algebra [18] that the objective functions 3 and 4 are equivalent.

Objective function 4 SIMPLS (Equivalent formulation)

$$(\boldsymbol{w}_i, \boldsymbol{u}_i) = \arg\max_{(\boldsymbol{w},\boldsymbol{u})} (\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{u})$$

subject to $\boldsymbol{w}_i^T \boldsymbol{w}_i = 1$, $\boldsymbol{u}_i^T \boldsymbol{u}_i = 1$ *and* $\boldsymbol{t}_i^T \boldsymbol{t}_j = \boldsymbol{w}_i^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}_j = 0$, *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` by Boulesteix and Strimmer [19].

A third -quite rarely used variant of PLS satisfying a simple objective function and denoted as undeflated PLS (UPLS) is proposed by Burnham et al. [20], rather as a piece of their global framework than to improve the prediction performance of the original PLS variants. The UPLS weight vectors satisfying objective function 5 are simply obtained as the eigenvectors of the matrix $\boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{X}$.

Objective function 5 UPLS

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} (\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{X}\boldsymbol{w})$$

subject to $\boldsymbol{w}_i^T \boldsymbol{w}_i = 1$ *and* $\boldsymbol{w}_i^T \boldsymbol{w}_j = 0$, *for* $j = 1, \ldots, i-1$.

It can be shown [16, 17] that the objective functions of PLS2 and SIMPLS are equivalent in the univariate case ($q = 1$). In the multivariate case, there is no rule of thumb as

to whether one should use PLS2 or SIMPLS. According to ter Braak and de Jong [16], it depends on the data. SIMPLS has become increasingly popular in the last decade because (i) of the simplicity of its criterion, (ii) of the computational efficiency of the algorithm(s), (iii) it can be seen as a generalization of the univariate case discussed in Section 2.3. In the literature, one can find many extensions of PLS regression that are reported to perform well in some situations. For example, Milidiu and Renteria [21] suggest two fast PLS procedures for very large data sets denoted as DPLS and PPLS, whereas Durand [22] proposes to perform PLS regression using splines transformations of the predictors to incorporate nonlinear structures.

2.5 Connections between PLS and OLS, PCR, CR, RR and RRR

Comparing PLS2 and SIMPLS to related methods such as Principal Component Regression (PCR) or Reduced Rank Regression (RRR) gives another interesting perspective on PLS dimension reduction/regression. In this section, we briefly review the connections of PLS to a few related approaches.

Objective function 6 PCR

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} (\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w})$$

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

The objective function of PCR looks quite similar to that of PLS, the only difference being that PCR maximizes the variance of the latent components, whereas PLS maximizes the covariance with the response. Stone and Brooks [12] are among the first to notice this similarity in the univariate case. They formulate the objective function of OLS, PLS and PCR as the unique objective function

$$(\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{Y}^T \boldsymbol{X} \boldsymbol{w})(\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w})^{\alpha/(1-\alpha)-1},$$

where α takes some value in the continuum $0 \leq \alpha \leq 1$. If $\alpha = 0$, the problem is equivalent to maximizing the sample correlation between Y and the latent component. The sequential construction terminates with $\boldsymbol{w}_1$ and the OLS regression coefficients are obtained. PLS1 is obtained for $\alpha = \frac{1}{2}$, whereas PCR is obtained for $\alpha \to 1$. The connection between Stone and Brook's Continuum Regression (CR) and standard Ridge Regression (RR) is studied by Sundberg [23]. It turns out that the vector of regression coefficients obtained with standard RR is proportional to the vector of regression coefficients from CR with one latent component, where the ridge parameter and CR parameter are monotonically related.

Another related dimension reduction technique is Reduced Rank Regression (RRR), which has the following objective function.

Objective function 7 RRR

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}}(\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{X}\boldsymbol{w})$$

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

The connection of RRR with PLS becomes more clear by considering the underlying model structure. The matrix $\boldsymbol{W}$ in Reduced Rank Regression is the least squares estimate of the model corresponding to equation (2) [24]:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{Q}^T + \boldsymbol{F}, \tag{5}$$

where $\boldsymbol{W}$ and $\boldsymbol{Q}$ have dimensions $p \times c$ and $q \times c$, respectively. Hence, RRR concentrates on the prediction and does not aim to explain the variation of the predictors. A maximum-likelihood interpretation of RRR is given in Burnham et al. [25]: RRR can be derived as the maximum-likelihood solution of model (5), if the rows of $\boldsymbol{F}$ are assumed i.i.d. multivariate normal with a known covariance matrix of the form $\sigma_Y 2\boldsymbol{I}_q$. In contrast to RRR, PCR can be seen as the least square solution of model (1) [24]:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}.$$

Thus, the latent components are constructed without consideration for the response matrix $\boldsymbol{Y}$, which often leads to poor prediction performance compared to methods making use of $\boldsymbol{Y}$. The weight vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c$ are the eigenvectors corresponding to the c largest eigen-values of $\boldsymbol{X}^T\boldsymbol{X}$.

In contrast, neither PLS2 nor SIMPLS can be seen as parameter estimations for a given model. Simulations performed by Burnham et al. [25] suggest that PLS might be somewhere between RRR and PCR when one considers a continuum regression method based on equations (1) and (2). In this framework, the rows of $\boldsymbol{E}$ and $\boldsymbol{F}$ are assumed have multivariate normal distributions $\mathcal{N}_p(0, \boldsymbol{\Sigma}_X)$ and $\mathcal{N}_q(0, \boldsymbol{\Sigma}_Y)$, respectively. If we have $\boldsymbol{\Sigma}_X = \sigma_X 2\boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_Y = \sigma_Y 2\boldsymbol{I}_q$ and $\phi = \sigma_X/\sigma_Y$, the maximum-likelihood solution leads to PCR if $\phi \to 0$ and to RRR if $\phi \to +\infty$. PLS is believed to lie somewhere in the middle, thus combining the advantages of RRR (supervised method) and PCR ($n < p$ is allowed).

3 Applications of PLS regression to high-dimensional microarray data

3.1 Regression

A straightforward application of univariate PLS regression to expression data from yeast *saccharomyces cerevisiae* can be found in Datta [26]. Some handpicked gene expression levels are regressed against expression levels of other genes using PLS univariate regression (PLS1) with different numbers of latent components. The magnitude of the obtained regression coefficients are interpreted in terms of interaction strength between genes. Unfortunately, in this paper many conclusions are drawn on a purely heuristic basis without concern for statistical relevance and validation. Furthermore, the choice of the number of latent components is purely heuristic.

Huang et al. [27] use PLS regression for another 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. [27] 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 coeffcients 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. [27] 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.

Applications of PLS multivariate regression include the prediction of transcription factor activities from combined analysis of gene expression data and ChIP data as proposed in Boulesteix and Strimmer [19]. 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 [19] assume as the relationship between microarray data and connectivity data the linear structure

$$\boldsymbol{Y} = \boldsymbol{A} + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

where $\boldsymbol{Y}$ is the $n \times q$ constant matrix containing the expression levels of n genes (rows) in

q conditions (columns), $\boldsymbol{X}$ is the $n \times p$ matrix containing the connectivity information for n genes (rows) and p transcription factors (columns), $\boldsymbol{A}$ is a $n \times q$ matrix corresponding to the intercepts and $\boldsymbol{E}$ is a $n \times q$ error matrix. The $p \times q$ matrix $\boldsymbol{B}$ 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 Boulesteix and Strimmer [19] by employing the SIMPLS method. Using PLS in this context allows not only to extract information on TFAs 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 [28]).

3.2 Classification

So far, we have considered only the case of continuous response variables. In many studies, however, the response to be predicted is categorical (either binary or multicategorical). Although PLS regression is designed for continuous response variables, it has often be applied with success to a categorical response. In the whole section, Y denotes a categorical response variable and $X_1, \ldots, X_p$ are the continuous predictors. In all the applications reviewed here, each of the n observations is a cancer patient, $X_1, \ldots, X_p$ are gene expression levels and Y is the tumor type of the considered patient.

It is important to distinguish binary response variables (with possible values $Y = 0, 1$) from multicategorical response variables (with possible values $Y = 0, \ldots, K-1$, where $K > 2$). Whereas binary variables may be treated as continuous variables in practice, this approach does not make sense with multicategorical (unordered) variables. If Y is multicategorical, it has to be transformed before PLS dimension reduction. A simple transformation method consists to convert Y into K random variables $Y_1, \ldots, Y_K$ defined as follows:

$$Y_j = \begin{cases} 1 & \text{if } Y = j-1 \\ 0 & \text{otherwise.} \end{cases}$$

In this framework, it can be shown that multivariate PLS dimension reduction almost leads to the same components as principal component analysis performed on the between-group sample covariance matrix. A collection of properties on this topic as well as mathematical proofs are given in Barker and Rayens [29]. These properties can be seen as a justification of PLS dimension reduction with categorical variables.

The most basic PLS-based approach to classification using a large number of predictors consists to treat Y (in the binary case) or $Y_1 \ldots, Y_K$ (in the multicategorical case) as

if they were continuous responses and to make the prediction by either univariate (in the binary case) or multivariate (in the multicategorical case) PLS regression. This approach is adopted by Huang and Pan [30] for binary response variables and compared to other statistical regression methods using the leukemia data by Golub et al. [31] and the colon cancer data by Alon et al. [32]. With this approach, one obtains continuous predictions generally ranging from about -1 to 2. Each observation is then assigned to one of the two classes 0 or 1, depending on the continuous prediction. Huang and Pan [30] suggest to determine the best number of latent components by leave-one-out cross-validation. Multivariate PLS regression is also employed in a more applied paper by Musumarra et al. [33] for the molecular diagnostic of cancer. Using the software `SIMCA`, they performed classification with the data set by Ross et al. [34] 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). This approach is reported to lead to high prediction accuracy, although it seems rather unappealing to predict categorical responses using a classical linear model.

Another related approach which is formally more correct consists to split the procedure into

1. a dimension reduction stage,

2. a classification stage consisting to apply a classical discrimination method (e.g. logistic regression, linear or quadratic discriminant analysis) using the PLS latent components as predictors.

The two-stage approach mentioned above is first proposed for a binary response by Nguyen and Rocke [1] and for a multicategorical response by Nguyen and Rocke [35], and further studied by Boulesteix [36]. To apply this method, one has to choose (i) the number of latent components to be extracted in the dimension reduction step, (ii) the classification method to be reduction for the classification step. In Nguyen and Rocke [35, 1], three classification methods are studied: logistic regression, linear discriminant analysis and quadratic discriminant analysis. For a general overview of these classification methods, see e.g. Hastie et al. [37]. Logistic regression turns out to be inappropriate because the maximum-likelihood estimate of the regression coefficients does not exist for separate and quasi-separate classes [38], which is a common situation when the PLS latent components are used as predictors. Linear discriminant analysis (LDA) turns out to yield the best classification performance, whereas quadratic discriminant analysis gives worse results. In Nguyen and Rocke [1] and Nguyen and Rocke [35], the number of PLS latent components is chosen on a heuristic basis as a 'tuning' parameter. In Boulesteix [36], the only investigated classification is linear discriminant analysis. The two-stage method

consisting of PLS dimension reduction and linear discriminant analysis will be denoted as PLS+LDA in the rest of the section. In Boulesteix [36] the choice of the number of latent components is addressed explicitly, thus making PLS+LDA a parameter-free classification method. This two-stage method (including the cross-validation procedure) is implemented in the `R` package `plsgenomics`. In the extensive comparison study performed by Boulesteix [36] including most state-of-art methods, PLS+LDA turns out to range among the best classification procedures for all the eight studied cancer data sets. Moreover, it can serve as a visualization tool to represent high-dimensional data in low dimension and can be connected to gene selection (see Section 3.3 for more details).

Although PLS dimension reduction/regression for categorical response variables may be interpreted in terms of an eigenvalue problem for the between-group covariance matrix [29], this approach has been criticized for its lack of formal coherence. Several attempts have been made to handle the case of categorical responses either by modifying the PLS regression algorithm or by using it as a piece of a more complicated classification procedure within the framework of generalized linear models. These PLS-based methods that are especially designed for the prediction of categorical variables are reviewed in Section 4.

3.3 Feature selection

An issue which 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 univariate PLS (PLS1) dimension reduction (see Section 2.3) applied to binary classification problems (see Section 3.2), the weight vector $\boldsymbol{w}_1 = (w_{11}, \ldots, w_{p1})^T$ defining the first latent component may be used to order the p genes in terms of their relevance for the classification problem [36]. Let F_j denote the F-statistic used in analysis of variance and computed from $\boldsymbol{X}$ for gene j as:

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

where

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

and

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

with n_k denoting the number of realizations 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 [36] 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 $\boldsymbol{X}$ 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 are more integrated 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. [39, 33]. It is based on all the weight vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c$ and implemented in the software package `SIMCA`. The 'variable influence' $\text{VIN}_{\gamma j}$ of gene j for the γ-th PLS component is defined as a function of $w_{j\gamma}^2$ 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' VIP_j, which is defined for each gene j as the sum of the $\text{VIN}_{\gamma j}$ 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.

3.4 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 $\{t, \delta, \boldsymbol{x}\}$, 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 the end of the study, and 0 if the patient was still alive at the end of the study,

- $\boldsymbol{x} = (X_1, \ldots, X_p)^T$ 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 the proportional hazard regression model (PH model) by Cox [40] 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 [41] suggest a two-stage method that consists to (i) perform univariate PLS with the failure time as response variable and $X_1, \ldots, X_p$ as predictors, (ii) use the obtained first latent components as predictors in classical PH regression. They apply their approach to a data set by [42] giving the survival time and expression levels of 5622 genes for 40 lymphoma patients, and to a data set by [43] 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 Nguyen [44], Park et al. [45] and Li and Gui [46]. Both approaches combine the construction of the successive PLS latent components with PH regression, but in different ways. They are reviewed in Section 4 which deals with PLS-based methods for special response variables.

4 PLS-based methods for special types of response variables

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 might seem unappealing, although it is reported to give good results in many cases. In this section, we review methods that use the principle of partial least squares 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. $\boldsymbol{Y}$ is a $n \times 1$ matrix.

4.1 Modification of the latent components in PLS regression

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

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

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

$$\begin{aligned} t_{i1} &\propto \sum_{j=1}^{p} \widehat{\mathrm{Cov}}(Y, X_j)x_{ij}, \\ &\propto \sum_{j=1}^{p} \widehat{\mathrm{Var}}(X_j)\hat{b}_j x_{ij}, \end{aligned}$$

where $\hat{b}_j$ 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 papers are based on the idea that $\hat{b}_j$ is not an optimal choice when Y is a binary or survival variable. Li and Gui [46] suggest to replace $\hat{b}_j$ 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. The idea consisting to replace a linear regression coefficient by a Cox regression coefficient also inspired another method denoted as "MPLS": Nguyen [44] gives a different

non-sequential expression of the PLS1 latent components $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_c$ involving eigenvectors of the matrices $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^T$ (see Nguyen and Rocke [47] for details). This complex expression also contains a linear regression coefficient, which Nguyen [44] replaces by a Cox regression coefficient.

The same approach is also used in the context of binary classification by Nguyen and Rocke [47] and denoted as "PLSM2". Another related PLS variant aiming to handle binary responses is also introduced in Nguyen and Rocke [47] under the name "PLSM1".

4.2 PLS and generalized linear models

Marx's IRPLS algorithm

Marx [48] 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 [49] for generalized linear models, resulting into 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 $\boldsymbol{X}$ is full row-rank, which is most often the case in high-dimensional microarray data with $n \ll p$ [50]. The IRPLS method as well as a few adaptations overcoming the convergence problem have been applied both to survival analysis and classification.

Application to 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 [51] and can be seen as an adaptation of Marx's IRPLS method which solves the problem of *separation*. As already mentioned in Section 3.2, infinite parameter estimates can occur in binary logistic regression when the two classes are completely or quasi-completely separated [38]. Firth [52] suggests a procedure to remove the first-order term of the asymptotic bias of maximum likelihood estimates in GLMs. The procedure is based on a modified score function which, when applied to logistic regression, guarantees finite estimates [53]. 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 Gentle-

man [51]. 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 by Alon et al. [32] and on the cancer data by Ross et al. [34]. The second approach to modify Marx's IRPLS is suggested by Fort and Lambert-Lacroix [50]: 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 by [32], the leukemia data by [31] and the prostate cancer data by [54].

Application to survival analysis

Another classical application of generalized linear models and IRPLS is survival analysis. As suggested by Whitehead [55], Park et al. [45] transform the failure time problem into a generalized linear regression problem with logarithmic link function. They propose to use the Iteratively Reweighted Partial Least Squares (IRPLS) estimation method for generalized linear regression described in Marx [48]. In contrast to the two-stage scheme developed in Nguyen and Rocke [41], 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 by [56]. According to Park et al. [45], 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 [46].

5 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 paper we have reviewed the partial least squares (PLS) approach to regression and dimension reduction that is perfectly suited for this kind of data. In particular, PLS automatically performs variable selection and can be applied to a diverse set of tasks, including classification, survival analysis, and modeling genetic networks.

We finally remark that in this review we have exclusively focused on applications of the PLS method to gene expression data. However, with the advent of proteomics data, e.g., from mass spectronometric experiments, we expect PLS to be further established as one of the prime tools for analyzing extremely high-dimensional molecular data.

A List of abbreviations

Term	Signification	Introduced in
PLS1	Univariate PLS	2.3
PLS2	Multivariate PLS (first)	2.4
SIMPLS	Multivariate PLS (second)	2.4
ULS	Undeflated PLS	2.4
OLS	Ordinary Least Squares	
PCR	Principal Component Analysis	2.5
RRR	Reduced Rank Regression	2.5
CR	Continuum Regression	2.5
RR	Ridge Regression	2.5
PLS+LDA	Two-step classification procedure consisting of PLS dimension reduction and LDA	3.2
IRPLS	Marx's Iteratively Reweighted PLS	4.2
$\boldsymbol{X} = (x_{ij})_{i=1,\dots,n,j=1,\dots,p}$	$n \times p$ matrix of predictors	2.2
$\boldsymbol{Y} = (y_{ij})_{i=1,\dots,n,j=1,\dots,p}$	$n \times q$ response matrix	2.2
$X_1, \dots, X_p$	Uncentered predictor variables (random variables)	2.2
$Y_1, \dots, Y_q$	Uncentered response variables	2.2
$(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{y}}_i)_{i=1,\dots,n}$	Uncentered sample	2.2
$(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1,\dots,n}$	Centered sample	2.2
$\boldsymbol{w}_j = (w_{1j}, \dots, w_{pj})^T$	weight vector defining the j-th latent component	2.2
$\boldsymbol{t}_j = (t_{1j}, \dots, t_{nj})^T$	j-th latent component	2.2
$\boldsymbol{T} = [\boldsymbol{t}_1, \dots, \boldsymbol{t}_c]$	$n \times c$ matrix of latent components	2.2
$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_c]$	$p \times c$ matrix of weights	2.2
$T_j,\ j = 1, \dots, c$	(Uncentered) random variable corresponding to $\boldsymbol{t}_j$	2.2
$\boldsymbol{P}$	$p \times c$ matrix of X-loadings	2.2
$\boldsymbol{Q}$	$q \times c$ matrix of Y-loadings	2.2
$\boldsymbol{E}$	$n \times p$ error matrix	2.2
$\boldsymbol{F}$	$n \times q$ error matrix	2.2
$\boldsymbol{B}$	$p \times q$ matrix of regression coefficients	2.2

B 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 mentioned in Section 3.2 and discussed by [1, 36].

- `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 Principal Component Regression (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 mentioned in Section 4.2 and proposed by [51].

- `plss`
  `http://www.math.univ-montp2.fr/~durand/ProgramSources.html`
  This package implements PLS regression based on splines transformations of the predictors as described in [22].

Other software: Classification with PLS regression is implemented in the software tool `SIMCA`. Several PLS algorithms are also implemented in the procedure `PLS` in `SAS`.

References

[1] D. Nguyen and D. M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50, 2002.

[2] K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342, 1991.

[3] K. C. Li. On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87:1025–1039, 1992.

[4] R. D. Cook and S. Weisberg. Discussion of "sliced inverse regression" by K. C. Li. *Journal of the American Statistical Association*, 86:328–332, 1991.

[5] H. Martens. Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 58:85–95, 2001.

[6] S. Wold. Personal memories of the early PLS development. *Chemometrics and Intelligent Laboratory Systems*, 58:83–84, 2001.

[7] H. Wold. *Estimation of principal components and related models by iterative least squares, in: P. R. Krishnaiah (Ed.), Multivariate Analysis*. Academic Press, New york, 1966.

[8] H. Wold. *Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments, in: P. R. Krishnaiah (Ed.), Multivariate Analysis*. Academic Press, New york, 1973.

[9] H. Wold. *Path models with latent variables: the NIPALS approach, in: H. M. Blalock (Ed.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*. Academic Press, New york, 1975.

[10] H. Schneeweiss. Models with latent variables: LISREL versus PLS. *Statistica Neerlandica*, 45:145–157, 1991.

[11] I. Helland. On the structure of partial least squares. *Communication in Statistics., Simulation and Computation*, 17:581–607, 1988.

[12] M. Stone and R. J. Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *Journal of the Royal Statistical Society B*, 52:237–269, 1990.

[13] I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135, 1993.

[14] P. H. Garthwaite. An interpretation of partial least squares. *Journal of the American Statistical Association*, 89:122–127, 1994.

[15] H. Martens and T. Naes. *Multivariate Calibration*. Wiley, New York, 1989.

[16] C. J. F. ter Braak and S. de Jong. The objective function of partial least squares regression. *Journal of Chemometrics*, 12:41–54, 1998.

[17] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–253, 1993.

[18] C. R. Rao. *Linear Statistical Inference and its Application*. Wiley, New York, 1973.

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

[20] A. J. Burnham, R. Viveros, and J. F. MacGregor. Frameworks for latent variable multivariate regression. *Journal of Chemometrics*, 10:31–45, 1996.

[21] R. L. Milidiu and R. P. Renteria. DPLS and PPLS: two algorithms for large data sets. *Computational Statistics and Data Analysis*, 48:125–138, 2005.

[22] J. F. Durand. Local polynomial additive regression through PLS and splines: PLSS. *Chemometrics and Intelligent Laboratory Systems*, 58:235–246, 2001.

[23] R. Sundberg. Continuum regression and ridge regression. *Journal of the Royal Statistical Society B*, 55:653–659, 1993.

[24] B. Abraham and G. Merola. Dimensionality reduction approach to multivariate prediction. *Computational Statistics and Data Analysis*, 48:5–16, 2005.

[25] A. J. Burnham, J. F. MacGregor, and R. Viveros. A statistical framework for latent variable multivariate regression methods based on maximum likelihood. *Journal of Chemometrics*, 13:49–65, 1999.

[26] S. Datta. Exploring relationships in gene expressions: a partial least squares approach. *Gene expression*, 9:257–264, 2001.

[27] X. Huang, W. Pan, S. Park, X. Han, L. W. Miller, and J. Hall. Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. *Journal of Chemometrics*, 20:888–894, 2004.

[28] M. Clementi, S. Clementi, G. Cruciani, M. Pastor, A. M. Davis, and D. R. Flower. Robust multivariate statistics and the prediction of protein secondary structure content. *Protein Engineering*, 10:747–749, 1997.

[29] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17:166–173, 2003.

[30] X. Huang and W. Pan. Linear regression and two-class classification with gene expression data. *Bioinformatics*, 19:2072–2078, 2003.

[31] T.R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J.R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[32] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96:6745–6750, 1999.

[33] G. Musumarra, V. Barresi, D. F. Condorelli, C. G. Fortuna, and S. Scire. Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis. *Journal of Chemometrics*, 18:125–132, 2004.

[34] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24:227–234, 2000.

[35] D. Nguyen and D. M. Rocke. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18:1216–1226, 2002.

[36] A. L. Boulesteix. PLS dimension reduction for classification with high-dimensional microarray data. *Statistical Applications in Genetics and Molecular Biology*, 3:Issue 3, Article 33, 2004.

[37] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning*. Springer-Verlag, New York, 2001.

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

[39] G. Musumarra, V. Barresi, D. F. Condorelli, and S. Scire. A bioinformatics approach to the identification of candidate genes for the development of new cancer diagnostics. *Biological Chemistry*, 384:321–327, 2003.

[40] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, B*, 34:187–220, 1972.

[41] D. Nguyen and D. M. Rocke. Partial least squares proportional hazard regression for application to dna microarray survival data. *Bioinformatics*, 18:1625–1632, 2002.

[42] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[43] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsten, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98:10869–10874, 2001.

[44] D. Nguyen. Partial least squares dimension reduction for microarray gene expression data with a censored response. *Mathematical Biosciences*, 193:119–137, 2005.

[45] P. J. Park, L. Tian, and I. S. Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18:120–127, 2002.

[46] H. Li and J. Gui. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20:208–215, 2004.

[47] D. V. Nguyen and D. M. Rocke. On partial least squares dimension reduction for microarray-based classification: a simulation study. *Computational Statistics and Data Analysis*, 46:407–425, 2004.

[48] B. D. Marx. Iteratively reweighted partial least squares. *Technometrics*, 38:374–381, 1996.

[49] P. Green. Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of the Royal Statistical Society, B*, 46:149–192, 1984.

[50] G. Fort and S. Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21:1104–1111, 2005.

[51] B. Ding and R. Gentleman. Classification using generalized partial least squares. *Bioconductor Project Working Papers*, page Paper 5, 2004.

[52] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

[53] G. Heinze and M. Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21:2409–2419, 2002.

[54] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

[55] J. Whitehead. Fitting Cox's regression model to survival data using GLIM. *Applied Statistics*, 29:268–275, 1980.

[56] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98: 13790–13795, 2001.