Overview and Recent Advances in Partial Least Squares

Roman Rosipal[1] and Nicole Krämer[2]

[1] Austrian Research Institute for Artificial Intelligence,
Freyung 6/6, A-1010 Vienna, Austria
roman.rosipal@ofai.at
[2] TU Berlin, Department of Computer Science and Electrical Engineering,
Franklinstraße 28/29, D-10587 Berlin, Germany
nkraemer@cs.tu-berlin.de

1 Introduction

Partial Least Squares (PLS) is a wide class of methods for modeling relations between sets of observed variables by means of latent variables. It comprises of regression and classification tasks as well as dimension reduction techniques and modeling tools. The underlying assumption of all PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent (not directly observed or measured) variables. Projections of the observed data to its latent structure by means of PLS was developed by Herman Wold and coworkers [48, 49, 52].

PLS has received a great amount of attention in the field of chemometrics. The algorithm has become a standard tool for processing a wide spectrum of chemical data problems. The success of PLS in chemometrics resulted in a lot of applications in other scientific areas including bioinformatics, food research, medicine, pharmacology, social sciences, physiology–to name but a few [28, 25, 53, 29, 18, 22].

This chapter introduces the main concepts of PLS and provides an overview of its application to different data analysis problems. Our aim is to present a concise introduction, that is, a valuable guide for anyone who is concerned with data analysis.

In its general form PLS creates orthogonal score vectors (also called latent vectors or components) by maximising the covariance between different sets of variables. PLS dealing with two blocks of variables is considered in this chapter, although the PLS extensions to model relations among a higher number of sets exist [44, 46, 47, 48, 39]. PLS is similar to Canonical Correlation Analysis (CCA) where latent vectors with maximal correlation are extracted [24]. There are different PLS techniques to extract latent vectors, and each of them gives rise to a variant of PLS.

PLS can be naturally extended to regression problems. The predictor and predicted (response) variables are each considered as a block of variables. PLS then extracts the score vectors which serve as a new predictor representation

C. Saunders et al. (Eds.): SLSFS 2005, LNCS 3940, pp. 34–51, 2006.

and regresses the response variables on these new predictors. The natural asymmetry between predictor and response variables is reflected in the way in which score vectors are computed. This variant is known under the names of PLS1 (one response variable) and PLS2 (at least two response variables). PLS regression used to be overlooked by statisticians and is still considered rather an algorithm than a rigorous statistical model [14]. Yet within the last years, interest in the statistical properties of PLS has risen. PLS has been related to other regression methods like Principal Component Regression (PCR) [26] and Ridge Regression (RR) [16] and all these methods can be cast under a unifying approach called continuum regression [40, 9]. The effectiveness of PLS has been studied theoretically in terms of its variance [32] and its shrinkage properties [12, 21, 7]. The performance of PLS is investigated in several simulation studies [11, 1].

PLS can also be applied to classification problems by encoding the class membership in an appropriate indicator matrix. There is a close connection of PLS for classification to Fisher Discriminant Analysis (FDA) [4]. PLS can be applied as a discrimination tool and dimension reduction method–similar to Principal Component Analysis (PCA). After relevant latent vectors are extracted, an appropriate classifier can be applied. The combination of PLS with Support Vector Machines (SVM) has been studied in [35].

Finally, the powerful machinery of kernel-based learning can be applied to PLS. Kernel methods are an elegant way of extending linear data analysis tools to nonlinear problems [38].

2 Partial Least Squares

Consider the general setting of a linear PLS algorithm to model the relation between two data sets (blocks of variables). Denote by $\mathcal{X} \subset \mathcal{R}^N$ an N-dimensional space of variables representing the first block and similarly by $\mathcal{Y} \subset \mathcal{R}^M$ a space representing the second block of variables. PLS models the relations between these two blocks by means of score vectors. After observing n data samples from each block of variables, PLS decomposes the $(n \times N)$ matrix of zero-mean variables $\mathbf{X}$ and the $(n \times M)$ matrix of zero-mean variables $\mathbf{Y}$ into the form

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{F} \end{aligned} \tag{1}$$

where the $\mathbf{T}$, $\mathbf{U}$ are $(n \times p)$ matrices of the p extracted score vectors (components, latent vectors), the $(N \times p)$ matrix $\mathbf{P}$ and the $(M \times p)$ matrix $\mathbf{Q}$ represent matrices of loadings and the $(n \times N)$ matrix $\mathbf{E}$ and the $(n \times M)$ matrix $\mathbf{F}$ are the matrices of residuals. The PLS method, which in its classical form is based on the nonlinear iterative partial least squares (NIPALS) algorithm [47], finds weight vectors $\mathbf{w}, \mathbf{c}$ such that

$$[cov(\mathbf{t}, \mathbf{u})]^2 = [cov(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2 = max_{|\mathbf{r}|=|\mathbf{s}|=1}[cov(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})]^2 \tag{2}$$

where $cov(\mathbf{t}, \mathbf{u}) = \mathbf{t}^T\mathbf{u}/n$ denotes the sample covariance between the score vectors $\mathbf{t}$ and $\mathbf{u}$. The NIPALS algorithm starts with random initialisation of the

$\mathcal{Y}$-space score vector $\mathbf{u}$ and repeats a sequence of the following steps until convergence.

1) $\mathbf{w} = \mathbf{X}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$	4) $\mathbf{c} = \mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$
2) $\|\mathbf{w}\| \to 1$	5) $\|\mathbf{c}\| \to 1$
3) $\mathbf{t} = \mathbf{X}\mathbf{w}$	6) $\mathbf{u} = \mathbf{Y}\mathbf{c}$

Note that $\mathbf{u} = \mathbf{y}$ if $M = 1$, that is, $\mathbf{Y}$ is a one-dimensional vector that we denote by $\mathbf{y}$. In this case the NIPALS procedure converges in a single iteration.

It can be shown that the weight vector $\mathbf{w}$ also corresponds to the first eigenvector of the following eigenvalue problem [17]

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \tag{3}$$

The $\mathcal{X}$- and $\mathcal{Y}$-space score vectors $\mathbf{t}$ and $\mathbf{u}$ are then given as

$$\mathbf{t} = \mathbf{X}\mathbf{w} \quad \text{and} \quad \mathbf{u} = \mathbf{Y}\mathbf{c} \tag{4}$$

where the weight vector $\mathbf{c}$ is define in steps 4 and 5 of NIPALS. Similarly, eigenvalue problems for the extraction of $\mathbf{t}$, $\mathbf{u}$ or $\mathbf{c}$ estimates can be derived [17]. The user then solves for one of these eigenvalue problems and the other score or weight vectors are readily computable using the relations defined in NIPALS.

2.1 Forms of PLS

PLS is an iterative process. After the extraction of the score vectors $\mathbf{t}$, $\mathbf{u}$ the matrices $\mathbf{X}$ and $\mathbf{Y}$ are deflated by subtracting their rank-one approximations based on $\mathbf{t}$ and $\mathbf{u}$. Different forms of deflation define several variants of PLS.

Using equations (1) the vectors of loadings $\mathbf{p}$ and $\mathbf{q}$ are computed as coefficients of regressing $\mathbf{X}$ on $\mathbf{t}$ and $\mathbf{Y}$ on $\mathbf{u}$, respectively

$$\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t}) \quad \text{and} \quad \mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$$

PLS Mode A: The PLS Mode A is based on rank-one deflation of individual block matrices using the corresponding score and loading vectors. In each iteration of PLS Mode A the $\mathbf{X}$ and $\mathbf{Y}$ matrices are deflated

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad \text{and} \quad \mathbf{Y} = \mathbf{Y} - \mathbf{u}\mathbf{q}^T$$

This approach was originally designed by Herman Wold [47] to model the relations between the different sets (blocks) of data. In contrast to the PLS regression approach, discussed next, the relation between the two blocks is symmetric. As such this approach seems to be appropriate for modeling existing relations between sets of variables in contrast to prediction purposes. In this way PLS Mode A is similar to CCA. Wegelin [45] discusses and compares properties of both methods. The connection between PLS and CCA from the point of an optimisation criterion involved in each method is discussed in Section 2.2.

PLS1, PLS2: PLS1 (one of the block of data consists of a single variable) and PLS2 (both blocks are multidimensional) are used as PLS regression methods. These variants of PLS are the most frequently used PLS approaches. The relationship between $\mathbf{X}$ and $\mathbf{Y}$ is asymmetric. Two assumptions are made: i) the score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ are good predictors of $\mathbf{Y}$; p denotes the number of extracted score vectors–PLS iterations ii) a linear inner relation between the scores vectors $\mathbf{t}$ and $\mathbf{u}$ exists; that is,

$$\mathbf{U} = \mathbf{TD} + \mathbf{H} \tag{5}$$

where $\mathbf{D}$ is the $(p \times p)$ diagonal matrix and $\mathbf{H}$ denotes the matrix of residuals. The asymmetric assumption of the predictor–predicted variable(s) relation is transformed into a deflation scheme where the predictor space, say $\mathbf{X}$, score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ are good predictors of $\mathbf{Y}$. The score vectors are then used to deflate $\mathbf{Y}$, that is, a component of the regression of $\mathbf{Y}$ on $\mathbf{t}$ is removed from $\mathbf{Y}$ at each iteration of PLS

$$\mathbf{X} = \mathbf{X} - \mathbf{tp}^T \quad \text{and} \quad \mathbf{Y} = \mathbf{Y} - \mathbf{tt}^T\mathbf{Y}/(\mathbf{t}^T\mathbf{t}) = \mathbf{Y} - \mathbf{tc}^T$$

where we consider not scaled to unit norm weight vectors $\mathbf{c}$ defined in step 4 of NIPALS. This deflation scheme guarantees mutual orthogonality of the extracted score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ [17]. Note that in PLS1 the deflation of $\mathbf{y}$ is technically not needed during the iterations of PLS [17].
Singular values of the cross-product matrix $\mathbf{X}^T\mathbf{Y}$ correspond to the sample covariance values [17]. Then the deflation scheme of extracting one component at a time has also the following interesting property. The first singular value of the deflated cross-product matrix $\mathbf{X}^T\mathbf{Y}$ at iteration $i+1$ is greater or equal than the second singular value of $\mathbf{X}^T\mathbf{Y}$ at iteration i [17]. This result can be also applied to the relation of eigenvalues of (3) due to the fact that (3) corresponds to the singular value decomposition of the transposed cross-product matrix $\mathbf{X}^T\mathbf{Y}$. In particular, the PLS1 and PLS2 algorithms differ from the computation of all eigenvectors of (3) in one step.

PLS-SB: As outlined at the end of the previous paragraph the computation of all eigenvectors of (3) at once would define another form of PLS. This computation involves a sequence of implicit rank-one deflations of the overall cross-product matrix. This form of PLS was used in [36] and in accordance with [45] it is denoted as PLS-SB. In contrast to PLS1 and PLS2, the extracted score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ are in general not mutually orthogonal.

SIMPLS: To avoid deflation steps at each iteration of PLS1 and PLS2, de Jong [8] has introduced another form of PLS denoted SIMPLS. The SIMPLS approach directly finds the weight vectors $\{\tilde{\mathbf{w}}\}_{i=1}^{p}$ which are applied to the original not deflated matrix $\mathbf{X}$. The criterion of the mutually orthogonal score vectors $\{\tilde{\mathbf{t}}\}_{i=1}^{p}$ is kept. It has been shown that SIMPLS is equal to PLS1 but differs from PLS2 when applied to the multidimensional matrix $\mathbf{Y}$ [8].

2.2 PCA, CCA and PLS

There exists a variety of different projection methods to latent variables. Among others widely used, PCA and CCA belong to this category. The connections between PCA, CCA and PLS can be seen through the optimisation criterion they use to define projection directions. PCA projects the original variables onto a direction of maximal variance called principal direction. Following the notation of (2), the optimisation criterion of PCA can be written as

$$max_{|\mathbf{r}|=1}[var(\mathbf{Xr})]$$

where $var(\mathbf{t}) = \mathbf{t}^T\mathbf{t}/n$ denotes the sample variance. Similarly CCA finds the direction of maximal correlation solving the following optimisation problem

$$max_{|\mathbf{r}|=|\mathbf{s}|=1}[corr(\mathbf{Xr}, \mathbf{Ys})]^2$$

where $[corr(\mathbf{t}, \mathbf{u})]^2 = [cov(\mathbf{t}, \mathbf{u})]^2/var(\mathbf{t})var(\mathbf{u})$ denotes the sample squared correlation. It is easy to see that the PLS criterion (2)

$$max_{|\mathbf{r}|=|\mathbf{s}|=1}[cov(\mathbf{Xr}, \mathbf{Ys})]^2 = max_{|\mathbf{r}|=|\mathbf{s}|=1}var(\mathbf{Xr})[corr(\mathbf{Xr}, \mathbf{Ys})]^2var(\mathbf{Ys}) \quad (6)$$

represents a form of CCA where the criterion of maximal correlation is balanced with the requirement to explain as much variance as possible in both $\mathcal{X}$- and $\mathcal{Y}$-spaces. Note that in the case of a one-dimensional $\mathcal{Y}$-space only the $\mathcal{X}$-space variance is involved.

The relation between CCA and PLS can be also seen through the concept of canonical ridge analysis introduced in [41]. Consider the following optimisation problem

$$\max_{|\mathbf{r}|=|\mathbf{s}|=1} \frac{cov(\mathbf{Xr}, \mathbf{Ys})^2}{([1-\gamma_{\mathbf{X}}]\, var(\mathbf{Xr}) + \gamma_{\mathbf{X}})\,([1-\gamma_{\mathbf{Y}}]\, var(\mathbf{Ys}) + \gamma_{\mathbf{Y}})}$$

with $0 \leq \gamma_{\mathbf{X}}, \gamma_{\mathbf{Y}} \leq 1$ representing regularisation terms. The corresponding eigenvalue problem providing the solution to this optimisation criterion is given as

$$([1-\gamma_{\mathbf{X}}]\mathbf{X}^T\mathbf{X} + \gamma_{\mathbf{X}}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}([1-\gamma_{\mathbf{Y}}]\mathbf{Y}^T\mathbf{Y} + \gamma_{\mathbf{Y}}\mathbf{I})^{-1}\mathbf{Y}^T\mathbf{Xw} = \lambda\mathbf{w} \quad (7)$$

where $\mathbf{w}$ represents a weight vector for the projection of the original $\mathcal{X}$-space data into a latent space.[1] There are two cornerstone solutions of this eigenvalue problem: i) for $\gamma_{\mathbf{X}} = 0, \gamma_{\mathbf{Y}} = 0$ the solution of CCA is obtained [24] ii) for $\gamma_{\mathbf{X}} = 1, \gamma_{\mathbf{Y}} = 1$ the PLS eigenvalue problem (3) is recovered. By continuous changing of $\gamma_{\mathbf{X}}, \gamma_{\mathbf{Y}}$ solutions lying between these two cornerstones are obtained. In Figure 1 the $\mathbf{w}$ directions for two-class problem as found by PLS, CCA and regularised CCA ($\gamma_{\mathbf{X}} = 0.99, \gamma_{\mathbf{Y}} = 0$) are plotted.

Another interesting setting is $\gamma_{\mathbf{X}} = 1, \gamma_{\mathbf{Y}} = 0$ which represents a form of orthonormalised PLS where the $\mathcal{Y}$-space data variance does not influence the

[1] In the analogous way the eigenvalue problem for the projections of the $\mathcal{Y}$-space data can be formulated.

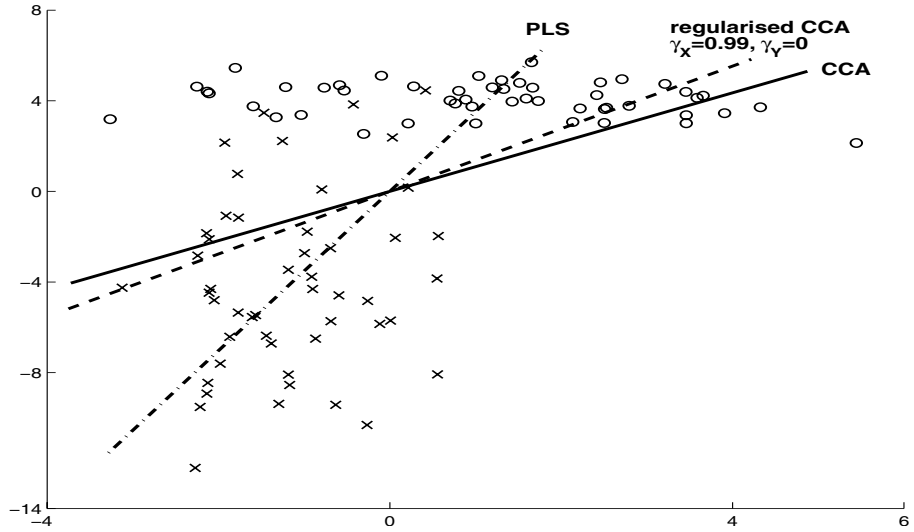

Fig. 1. An example of the weight vector $\mathbf{w}$ directions as found by CCA (solid line), PLS (dash-dotted line) and regularised CCA (dashed line) given by (7) with $\gamma_{\mathbf{X}} = 0.99$ and $\gamma_{\mathbf{Y}} = 0$. Circle and cross samples represent two Gaussian distributed classes with different sample means and covariances.

final PLS solution (similarly the $\mathcal{X}$-space variance can be ignored by setting $\gamma_{\mathbf{X}} = 0, \gamma_{\mathbf{Y}} = 1$) [53]. Note that in the case of one-dimensional $\mathbf{Y}$ matrix and for $\gamma_{\mathbf{X}} \in (0, 1)$ the ridge regression solution is obtained [41, 16]. Finally let us stress that, in general, CCA is solved in a way similar to PLS-SB, that is, eigenvectors and eigenvalues of (7) are extracted at once by an implicit deflation of the cross-product matrix $\mathbf{X}^T\mathbf{Y}$. This is in contrast to the PLS1 and PLS2 approaches where different deflation scheme is considered.

3 PLS Regression

As mentioned in the previous section, PLS1 and PLS2 can be used to solve linear regression problems. Combining assumption (5) of a linear relation between the scores vectors $\mathbf{t}$ and $\mathbf{u}$ with the decomposition of the $\mathbf{Y}$ matrix, equation (1) can be written as

$$\mathbf{Y} = \mathbf{TDQ}^T + (\mathbf{HQ}^T + \mathbf{F})$$

This defines the equation

$$\mathbf{Y} = \mathbf{TC}^T + \mathbf{F}^* \tag{8}$$

where $\mathbf{C}^T = \mathbf{DQ}^T$ now denotes the $(p \times M)$ matrix of regression coefficients and $\mathbf{F}^* = \mathbf{HQ}^T + \mathbf{F}$ is the residual matrix. Equation (8) is simply the decomposition of $\mathbf{Y}$ using ordinary least squares regression with orthogonal predictors $\mathbf{T}$.

We now consider orthonormalised score vectors $\mathbf{t}$, that is, $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, and the matrix $\mathbf{C} = \mathbf{Y}^T\mathbf{T}$ of the not scaled to length one weight vectors $\mathbf{c}$. It is useful

to redefine equation (8) in terms of the original predictors $\mathbf{X}$. To do this, we use the relationship [23]

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}$$

where $\mathbf{P}$ is the matrix of loading vectors defined in (1). Plugging this relation into (8), we yield

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{F}^*$$

where $\mathbf{B}$ represents the matrix of regression coefficients

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{C}^T = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y}$$

For the last equality, the relations among $\mathbf{T}$, $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{P}$ are used [23, 17, 33]. Note that different scalings of the individual score vectors $\mathbf{t}$ and $\mathbf{u}$ do not influence the $\mathbf{B}$ matrix. For training data the estimate of PLS regression is

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} = \mathbf{T}\mathbf{T}^T\mathbf{Y} = \mathbf{T}\mathbf{C}^T$$

and for testing data we have

$$\hat{\mathbf{Y}}_t = \mathbf{X}_t\mathbf{B} = \mathbf{T}_t\mathbf{T}^T\mathbf{Y} = \mathbf{T}_t\mathbf{C}^T$$

where $\mathbf{X}_t$ and $\mathbf{T}_t = \mathbf{X}_t\mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}$ represent the matrices of testing data and score vectors, respectively.

3.1 Algebraic Interpretation of Linear Regression

In this paragraph, we only consider PLS1, that is, the output data $\mathbf{y}$ is a one-dimensional vector. The linear regression model is usually subsumed in the relation

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{9}$$

with $\mathbf{b}$ the unknown regression vector and $\mathbf{e}$ a vector of independent identically distributed noise with $var(\mathbf{e}) = \sigma^2$. In what follows, we will make intensive use of the singular value decomposition of $\mathbf{X}$

$$\mathbf{X} = \mathbf{V}\mathbf{\Sigma}\mathbf{S}^T \tag{10}$$

with $\mathbf{V}$ and $\mathbf{S}$ orthonormal matrices and $\mathbf{\Sigma}$ a diagonal matrix that consists of the singular values of $\mathbf{X}$. The matrix $\mathbf{\Lambda} = \mathbf{\Sigma}^2$ is diagonal with elements λ_i . Set

$$\mathbf{A} \equiv \mathbf{X}^T\mathbf{X} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T \quad \text{and} \quad \mathbf{z} \equiv \mathbf{X}^T\mathbf{y}$$

The ordinary least squares (OLS) estimator $\hat{\mathbf{b}}_{OLS}$ is the solution of

$$arg\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

This problem is equivalent to computing the solution of the normal equations

$$\mathbf{A}\mathbf{b} = \mathbf{z} \tag{11}$$

Using the pseudoinverse of $\mathbf{A}^-$, it follows (recall (10)) that

$$\hat{\mathbf{b}}_{OLS} = \mathbf{A}^-\mathbf{z} = \sum_{i=1}^{rk(\mathbf{A})} \frac{\mathbf{v}_i^T\mathbf{y}}{\sqrt{\lambda_i}}\,\mathbf{s}_i = \sum_{i=1}^{rk(\mathbf{A})} \hat{\mathbf{b}}_i$$

where

$$\hat{\mathbf{b}}_i = \frac{\mathbf{v}_i^T\mathbf{y}}{\sqrt{\lambda_i}}\,\mathbf{s}_i$$

is the component of $\hat{\mathbf{b}}_{OLS}$ along $\mathbf{v}_i$ and $rk(.)$ denotes the rank of a matrix.

A lot of linear regression estimators are approximate solutions of the equation (11). The PCR estimator that regresses $\mathbf{y}$ on the first p principal components $\mathbf{v}_1, \ldots, \mathbf{v}_p$ is

$$\hat{\mathbf{b}}_{PCR} = \sum_{i=1}^{p} \hat{\mathbf{b}}_i$$

The RR estimator [41, 16] is of the form

$$\hat{\mathbf{b}}_{RR} = (\mathbf{A} + \gamma\mathbf{I})^{-1}\mathbf{z} = \sum_{i=1}^{rk(\mathbf{A})} \frac{\lambda_i}{\lambda_i + \gamma}\hat{\mathbf{b}}_i$$

with $\gamma > 0$ the ridge parameter.

It can be shown that the PLS algorithm is equivalent to the conjugate gradient method [15]. This is a procedure that iteratively computes approximate solutions of (11) by minimising the quadratic function

$$\frac{1}{2}\mathbf{b}^T\mathbf{A}\mathbf{b} - \mathbf{z}^T\mathbf{b}$$

along directions that are $\mathbf{A}$-orthogonal. The approximate solution obtained after p steps is equal to the PLS estimator obtained after p iterations.

The conjugate gradient algorithm is in turn closely related to the Lanczos algorithm [19], a method for approximating eigenvalues. The space spanned by the columns of

$$\mathbf{K} = (\mathbf{z}, \mathbf{A}\mathbf{z}, \ldots, \mathbf{A}^{p-1}\mathbf{z})$$

is called the p-dimensional Krylov space of $\mathbf{A}$ and $\mathbf{z}$. We denote this Krylov space by $\mathcal{K}$. In the Lanczos algorithm, an orthogonal basis

$$\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_p) \tag{12}$$

of $\mathcal{K}$ is computed. The linear map $\mathbf{A}$ restricted to $\mathcal{K}$ for an element $\mathbf{k} \in \mathcal{K}$ is defined as the orthogonal projection of $\mathbf{A}\mathbf{k}$ onto the space $\mathcal{K}$. The map is represented by the $p \times p$ matrix

$$\mathbf{L} = \mathbf{W}^T\mathbf{A}\mathbf{W}$$

This matrix is tridiagonal. Its p eigenvector-eigenvalue pairs

$$(\mathbf{r}_i\,,\,\mu_i) \tag{13}$$

are called Ritz pairs. They are the best approximation of the eigenpairs of $\mathbf{A}$ given only the information that is encoded in $\mathcal{K}$ [30].

The weight vectors $\mathbf{w}$ in (2) of PLS1 are identical to the basis vectors in (12). In particular, the weight vectors are a basis of the Krylov space and the PLS estimator is the solution of the optimisation problem

$$\begin{aligned} &arg\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \\ &subject\ to \quad \mathbf{b} \in \mathcal{K} \end{aligned}$$

In this sense, PLS1 can be viewed as a regularised least squares fit.

A good references for the Lanczos method and the conjugate gradient method is [30]. The connection to PLS is well-elaborated in [31].

3.2 Shrinkage Properties of PLS Regression

One possibility to evaluate the quality of an estimator $\hat{\mathbf{b}}$ for $\mathbf{b}$ is to determine its Mean Squared Error (MSE), which is defined as

$$\begin{aligned} \mathrm{MSE}(\hat{\mathbf{b}}) &= E\left[\left(\hat{\mathbf{b}} - \mathbf{b}\right)^T \left(\hat{\mathbf{b}} - \mathbf{b}\right)\right] \\ &= \left(E\left[\hat{\mathbf{b}}\right] - \mathbf{b}\right)^T \left(E\left[\hat{\mathbf{b}}\right] - \mathbf{b}\right) + E\left[\left(\hat{\mathbf{b}} - E\left[\hat{\mathbf{b}}\right]\right)^T \left(\hat{\mathbf{b}} - E\left[\hat{\mathbf{b}}\right]\right)\right] \end{aligned}$$

This is the well-known bias-variance decomposition of MSE. The first part is the squared bias and the second part is the variance term.

It is well known that the OLS estimator has no bias (if $\mathbf{b} \in range(\mathbf{A})$). The variance term depends on the non-zero eigenvalues of $\mathbf{A}$: if some eigenvalues are very small, the variance of $\hat{\mathbf{b}}_{OLS}$ can be very high, which leads to a high MSE value. Note that small eigenvalues λ_i of $\mathbf{A}$ correspond to principal directions $\mathbf{v}_i$ of $\mathbf{X}$ that have a low sample spread.

One possibility to decrease MSE is to modify the OLS estimator by shrinking the directions of the OLS estimator that are responsible for a high variance. In general, a shrinkage estimator for $\mathbf{b}$ is of the form

$$\hat{\mathbf{b}}_{shr} = \sum_{i=1}^{rk(\mathbf{A})} f(\lambda_i)\hat{\mathbf{b}}_i \tag{14}$$

where $f(.)$ is some real-valued function. The values $f(\lambda_i)$ are called shrinkage factors. Examples are PCR

$$f(\lambda_i) = \begin{cases} 1\ , & i \leq p \\ 0\ , & i > p \end{cases}$$

and RR

$$f(\lambda_i) = \frac{\lambda_i}{\lambda_i + \gamma}$$

If the factors in (14) do not depend on $\mathbf{y}$, that is, $\hat{\mathbf{b}}_{shr}$ is linear in $\mathbf{y}$, any factor $f(\lambda_i) \neq 1$ increases the bias of the i-th component. The variance of the i-th component decreases for $|f(\lambda_i)| < 1$ and increases for $|f(\lambda_i)| > 1$. The OLS estimator is shrunk in the hope that the increase in bias is small compared to the decrease in variance.

The PLS estimator is a shrinkage estimator as well. Its shrinkage factors are closely related to the Ritz pairs (13). The shrinkage factors $f(\lambda_i)$ that correspond to the estimator $\hat{\mathbf{b}}_{PLS}$ after p iterations of PLS are [21, 31]

$$f(\lambda_i) = 1 - \prod_{j=1}^{p} \left(1 - \frac{\lambda_i}{\mu_j}\right)$$

The shrinkage factors have some remarkable properties [7, 21]. Most importantly, $f(\lambda_i) > 1$ can occur for certain combinations of i and p. Note however that the PLS estimator is not linear in $\mathbf{y}$. The factors $f(\lambda_i)$ depend on the eigenvalues (13) of the matrix $\mathbf{L}$ and $\mathbf{L}$ in turn depends, via $\mathbf{z}$, on $\mathbf{y}$. It is therefore not clear in which way this shrinkage behaviour influences MSE of PLS1.

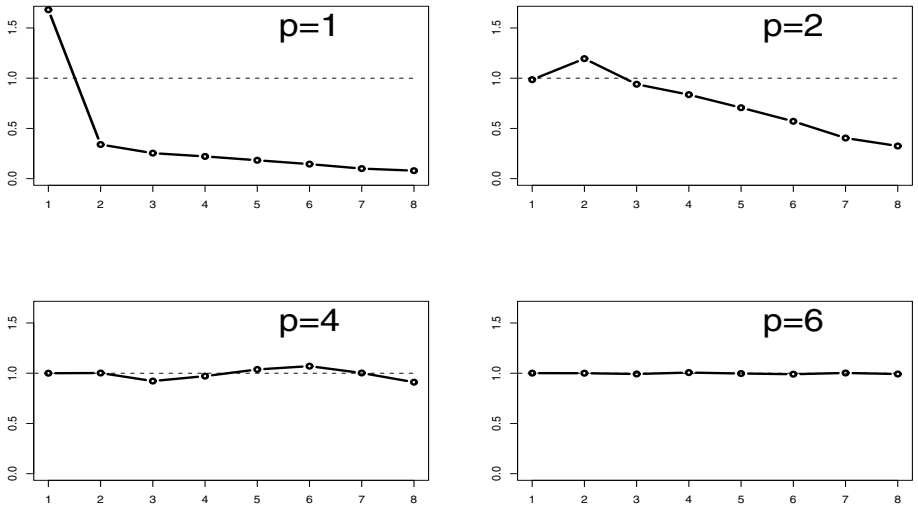

Fig. 2. An illustration of the shrinkage behaviour of PLS1. The $\mathbf{X}$ matrix contains eight variables. The eigenvalues of $\mathbf{A} \equiv \mathbf{X}^T\mathbf{X}$ are enumerated in decreasing order, $\lambda_1 \geq \lambda_2 \geq \ldots$ and the shrinkage factors $f(\lambda_i)$ are plotted as a function of i. The amount of absolute shrinkage $|1 - f(\lambda_i)|$ is particularly prominent if p is small.

4 PLS Discrimination and Classification

PLS has been used for discrimination and classification purposes. The close connection between FDA, CCA and PLS in the discrimination and classification scenario is described in this section.

Consider a set of n samples $\{\mathbf{x}_i \in \mathcal{X} \subset \mathcal{R}^N\}_{i=1}^n$ representing the data from g classes (groups). Now define the $(n \times g-1)$ class membership matrix $\mathbf{Y}$ to be

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \mathbf{1}_{n_{g-1}} \\ \mathbf{0}_{n_g} & \mathbf{0}_{n_g} & \dots & \mathbf{0}_{n_g} \end{pmatrix}$$

where $\{n_i\}_{i=1}^g$ denotes the number of samples in each class, $\sum_{i=1}^g n_i = n$ and $\mathbf{0}_{n_i}$ and $\mathbf{1}_{n_i}$ are $(n_i \times 1)$ vectors of all zeros and ones, respectively. Let

$$\mathbf{S_X} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \quad , \quad \mathbf{S_Y} = \frac{1}{n-1}\mathbf{Y}^T\mathbf{Y} \quad \text{and} \quad \mathbf{S_{XY}} = \frac{1}{n-1}\mathbf{X}^T\mathbf{Y}$$

be the sample estimates of the covariance matrices $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_Y}$, respectively, and the cross-product covariance matrix $\mathbf{\Sigma_{XY}}$. Again, the matrices $\mathbf{X}$ and $\mathbf{Y}$ are considered to be zero-mean. Furthermore, let

$$\mathbf{H} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \quad , \quad \mathbf{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_i^j - \bar{\mathbf{x}}_i)(\mathbf{x}_i^j - \bar{\mathbf{x}}_i)^T$$

represent the *among-classes* and *within-classes* sums-of-squares, where $\mathbf{x}_i^j$ represents an N-dimensional vector for the j-th sample in the i-th class and

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_i^j \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathbf{x}_i^j$$

Fisher developed a discrimination method based on a linear projection of the input data such that among-classes variance is maximised relative to the within-classes variance. The directions onto which the input data are projected are given by the eigenvectors $\mathbf{a}$ of the eigenvalue problem

$$\mathbf{E}^{-1}\mathbf{H}\mathbf{a} = \lambda\mathbf{a}$$

In the case of discriminating multi-normally distributed classes with the same covariance matrices, FDA finds the same discrimination directions as linear discriminant analysis using Bayes theorem to estimate posterior class probabilities. This is the method that provides the discrimination rule with minimal expected misclassification error [24, 13].

The fact that the Fisher's discrimination directions are identical to the directions given by CCA using a dummy matrix $\mathbf{Y}$ for group membership was first recognised in [5]. The connections between PLS and CCA have been methodically studied in [4]. Among other, the authors argue that the $\mathcal{Y}$-space penalty $var(\mathbf{Ys})$ is not meaningful and suggested to remove it from (6) in the PLS discrimination scenario. As mentioned in Section 2.2 this modification leads to a special case of the previously proposed orthonormalised PLS method [53] using the indicator

matrix $\mathbf{Y}$. The eigenvalue problem (3) in the case of orthonormalised PLS is transformed into

$$\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \tag{15}$$

where

$$\tilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$$

represents a matrix of uncorrelated and normalised output variables. Using the following relation [4, 35]

$$(n-1)\mathbf{S}_{\mathbf{XY}}\mathbf{S}_{\mathbf{Y}}^{-1}\mathbf{S}_{\mathbf{XY}}^T = \mathbf{H}$$

the eigenvectors of (15) are equivalent to the eigensolutions of

$$\mathbf{H}\mathbf{w} = \lambda\mathbf{w} \tag{16}$$

Thus, this modified PLS method is based on eigensolutions of the among-classes sum-of-squares matrix $\mathbf{H}$ which connects this approach to CCA or equivalently to FDA.

Interestingly, in the case of two-class discrimination the direction of the first orthonormalised PLS score vector $\mathbf{t}$ is identical with the first score vector found by either the PLS1 or PLS-SB methods. This immediately follows from the fact that $\mathbf{Y}^T\mathbf{Y}$ is a number in this case. In this two-class scenario $\mathbf{X}^T\mathbf{Y}$ is of a rank-one matrix and PLS-SB extracts only one score vector $\mathbf{t}$. In contrast, orthonormalised PLS can extract additional score vectors, up to the rank of $\mathbf{X}$, each being similar to directions computed with CCA or FDA on deflated feature space matrices. Thus, PLS provide more principled dimensionality reduction in comparison to PCA based on the criterion of maximum data variation in the $\mathcal{X}$-space alone.

In the case of multi-class discrimination the rank of the $\mathbf{Y}$ matrix is equal to $g-1$ which determines the maximum number of score vectors that may be extracted by the orthonormalised PLS-SB method.[2] Again, similar to the one-dimensional output scenario the deflation of the $\mathbf{Y}$ matrix at each step can be done using the score vectors $\mathbf{t}$ of PLS2. Consider this deflation scheme in the $\mathcal{X}$- and $\mathcal{Y}$-spaces

$$\mathbf{X}_d = \mathbf{X} - \mathbf{t}\mathbf{p}^T = (\mathbf{I} - \mathbf{t}\mathbf{t}^T/(\mathbf{t}^T\mathbf{t}))\mathbf{X} = \mathbf{P}_d\mathbf{X}$$
$$\tilde{\mathbf{Y}}_d = \mathbf{P}_d\tilde{\mathbf{Y}}$$

where $\mathbf{P}_d = \mathbf{P}_d^T\mathbf{P}_d$ represents a projection matrix. Using these deflated matrices $\mathbf{X}_d$ and $\tilde{\mathbf{Y}}_d$ the eigenproblem (15) can be written in the form

$$\mathbf{X}_d^T\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\mathbf{X}_d\mathbf{w} = \lambda\mathbf{w}$$

[2] It is considered here that $g \leq N$, otherwise the number of score vectors is given by N.

Thus, similar to the previous two-class discrimination the solution of this eigenproblem can be interpreted as the solution of (16) using the among-classes sum-of-squares matrix now computed on deflated matrix $\mathbf{X}_d$.

A natural further step is to project the original, observed data onto the obtained weight vector directions and to build a classifier using this new, projected data representation–PLS score vectors. Support vector machines, logistic regression or other methods for classification can be applied on the extracted PLS score vectors.

5 Nonlinear PLS

In many areas of research and industrial situations data can exhibit nonlinear behaviour. Two major approaches to model nonlinear data relations by means of PLS exist.

A) The first group of approaches is based on reformulating the considered linear relation (5) between the score vectors $\mathbf{t}$ and $\mathbf{u}$ by a nonlinear model

$$\mathbf{u} = g(\mathbf{t}) + \mathbf{h} = g(\mathbf{X}, \mathbf{w}) + \mathbf{h}$$

where $g(.)$ represents a continuous function modeling the existing nonlinear relation. Again, $\mathbf{h}$ denotes a vector of residuals. Polynomial functions, smoothing splines, artificial neural networks or radial basis function networks have been used to model $g(.)$ [51, 10, 50, 3].[3] The assumption that the score vectors $\mathbf{t}$ and $\mathbf{u}$ are linear projections of the original variables is kept. This leads to the necessity of a linearisation of the nonlinear mapping $g(.)$ by means of Taylor series expansions and to the successive iterative update of the weight vectors $\mathbf{w}$ [51, 3].

B) The second approach to nonlinear PLS is based on a mapping of the original data by means of a nonlinear function to a new representation (data space) where linear PLS is applied. The recently developed theory of kernel-based learning has been also applied to PLS. The nonlinear kernel PLS methodology was proposed for the modeling of relations between sets of observed variables, regression and classification problems [34, 35]. The idea of the kernel PLS approach is based on the mapping of the original $\mathcal{X}$-space data into a high-dimensional feature space $\mathcal{F}$ corresponding to a reproducing kernel Hilbert space [2, 38]

$$\mathbf{x} \in \mathcal{X} \rightarrow \mathbf{\Phi}(\mathbf{x}) \in \mathcal{F}$$

By applying the *kernel trick* the estimation of PLS in a feature space $\mathcal{F}$ reduces to the use of linear algebra as simple as in linear PLS [34]. The kernel trick uses the fact that a value of a dot product between two vectors in $\mathcal{F}$ can be evaluated by the kernel function [2, 38]

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x})^T \mathbf{\Phi}(\mathbf{y}), \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$$

[3] Note that the below described concept of kernel-based learning can also be used for modeling nonlinear relation between $\mathbf{t}$ and $\mathbf{u}$. An example would be a support vector regression model for $g(.)$ [38].

Define the Gram matrix $\mathbf{K}$ of the cross dot products between all mapped input data points, that is, $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$, where $\mathbf{\Phi}$ denotes the matrix of mapped $\mathcal{X}$-space data $\{\mathbf{\Phi}(\mathbf{x}_i) \in \mathcal{F}\}_{i=1}^n$. The kernel trick implies that the elements i, j of $\mathbf{K}$ are equal to the values of the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. Now, consider a modified version of the NIPALS algorithm where we merge steps 1 and 3 and we scale to unit norm vectors $\mathbf{t}$ and $\mathbf{u}$ instead of the vectors $\mathbf{w}$ and $\mathbf{c}$. We obtain the kernel form of the NIPALS algorithm [34, 20][4]

1) $\mathbf{t} = \mathbf{\Phi}\mathbf{\Phi}^T\mathbf{u} = \mathbf{K}\mathbf{u}$
2) $\|\mathbf{t}\| \to 1$
3) $\mathbf{c} = \mathbf{Y}^T\mathbf{t}$
4) $\mathbf{u} = \mathbf{Y}\mathbf{c}$
5) $\|\mathbf{u}\| \to 1$

Note that steps 3 and 4 can be further merged which may become useful in applications where an analogous kernel mapping of the $\mathcal{Y}$-space is considered. The kernel PLS approach has been proved to be competitive with the other kernel classification and regression approaches like SVM, kernel RR or kernel FDA [38, 37].

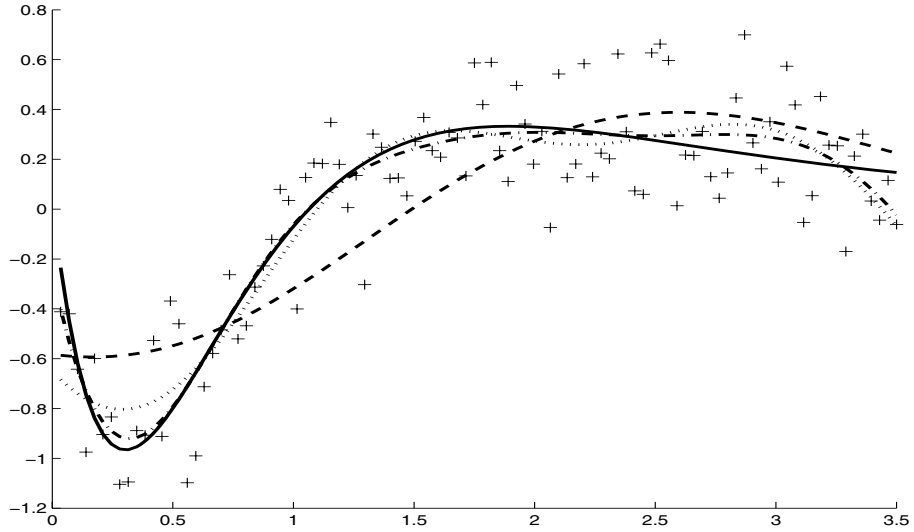

Fig. 3. An example of kernel PLS regression. The generated function $z(.)$ is shown as a solid line. Plus markers represent noisy representation of $z(.)$ used as training output points in kernel PLS regression. Kernel PLS regression using the first one, four and eight score vectors is shown as a dashed, dotted and dash-dotted line, respectively.

When both A) and B) approaches are compared it is difficult to define the favourable methodology. While the kernel PLS approach is easily implementable, computationally less demanding and capable to model difficult nonlinear relations, a loss of the interpretability of the results with respect to the original

[4] In the case of the one-dimensional $\mathcal{Y}$-space computationally more efficient kernel PLS algorithms have been proposed in [35, 27].

data limits its use in some applications. On the other hand it is not difficult to construct data situations where the first approach of keeping latent variables as linear projections of the original data may not be adequate. In practice a researcher needs to decide about the adequacy of using a particular approach based on the problem in hands and requirements like simplicity of the solution and implementation or interpretation of the results.

In Figure 3 an example of kernel PLS regression is depicted. We generated one hundred uniformly spaced samples in the range $[0, 3.5]$ and computed the corresponding values of the function [42]

$$z(x) = 4.26(\exp(-x) - 4\exp(-2x) + 3\exp(-3x))$$

Additional one hundred Gaussian distributed samples with zero-mean and variance equal to 0.04 representing noise were generated and added to the computed values. The values of noisy $z(.)$ function were subsequently centered. The Gaussian kernel function $k(\mathbf{x}, \mathbf{y}) = exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{h})$ with the width h equal to 1.8 was used.

6 Conclusions

PLS has been proven to be a very powerful versatile data analytical tool applied in many areas of research and industrial applications. Computational and implementation simplicity of PLS is a strong aspect of the approach which favours PLS to be used as a first step to understand the existing relations and to analyse real world data. The PLS method projects original data onto a more compact space of latent variables. Among many advantages of such an approach is the ability to analyse the importance of individual observed variables potentially leading to the deletion of unimportant variables. This mainly occurs in the case of an experimental design where many insignificant terms are measured. In such situations PLS can guide the practitioner into more compact experimental settings with a significant cost reduction and without a high risk associated with the "blind" variables deletion. Examples of this aspect of PLS are experiments on finger movement detection and cognitive fatigue prediction where a significant reduction of the EEG recording electrodes have been achieved without the loss of classification accuracy of the considered PLS models [35, 43]. Further important aspect of PLS is the ability to visualise high-dimensional data through the set of extracted latent variables. The diagnostic PLS tools based on score and loadings plots allows to better understand data structure, observe existing relations among data sets but also to detect outliers in the measured data.

Successful application of PLS on regression problems associated with many real world data have also attracted attention of statisticians to this method. Although PLS regression is still considered as a method or algorithm rather than a rigorous statistical model, recent advances in understanding of shrinkage properties of PLS regression helped to connect PLS regression with other, in statistical community better understood, shrinkage regression methods like PCR

or RR. Moreover, these studies have shown very competitive behaviour of PLS regression in comparison to the other shrinkage regression methods. We believe that further research will reveal additional aspects of PLS regression and will help to better theoretically define structures of data and regression problems where the use of PLS will become beneficial in comparison to the other methods.

Two major approaches of constructing nonlinear PLS have been mentioned. Among other nonlinear versions of PLS, kernel PLS represents an elegant way of dealing with nonlinear aspects of measured data. This method keeps computational and implementation simplicity of linear PLS while providing a powerful modeling, regression, discrimination or classification tool. Kernel PLS approach has been proven to be competitive with the other state-of-the-art kernel-based regression and classification methods.

Connections between PCA, (regularised) CCA and PLS have been highlighted (see [6] for detailed comparison). Understanding of these connections should help to design new algorithms by combining good properties of individual methods and thus resulting in more powerful machine learning tools.

References

1. T. Almøy. A simulation study on comparison of prediction models when only a few components are relevant. *Computational Statistics and Data Analysis*, 21:87–107, 1996.
2. N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.
3. G. Baffi, E.B. Martin, and A.J. Morris. Non-linear projection to latent structures revisited (the neural network PLS algorithm). *Computers Chemical Engineering*, 23:1293–1307, 1999.
4. M. Barker and W.S. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17:166–173, 2003.
5. M.S. Bartlett. Further aspects of the theory of multiple regression. In *Proceedings of the Cambridge Philosophical Society*, volume 34, pages 33–40, 1938.
6. T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in Pattern Recognition. In E. Bayro-Corrochano, editor, *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics*, pages 129–170. Springer, 2005.
7. N.A. Butler and M.C. Denham. The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society: B*, 62:585–593, 2000.
8. S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993.
9. S. de Jong, B.M. Wise, and N.L. Ricker. Canonical partial least squares and continuum power regression. *Journal of Chemometrics*, 15:85–100, 2001.
10. I.E. Frank. A nonlinear PLS model. *Chemolab*, 8:109–119, 1990.
11. I.E. Frank and J.H. Friedman. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35:109–147, 1993.
12. C. Goutis. Partial least squares yields shrinkage estimators. *The Annals of Statistics*, 24:816–824, 1996.
13. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

14. I.S. Helland. Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58:97–107, 1999.
15. M. Hestenes and E. Stiefel. Methods for conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.
16. A.E. Hoerl and R.W. Kennard. Ridge regression: bias estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.
17. A. Höskuldsson. PLS Regression Methods. *Journal of Chemometrics*, 2:211–228, 1988.
18. J.S. Hulland. Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic Management Journal*, 20:195–204, 1999.
19. C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45:225–280, 1950.
20. P.J. Lewi. Pattern recognition, reflection from a chemometric point of view. *Chemometrics and Intelligent Laboratory Systems*, 28:23–33, 1995.
21. O.C. Lingjærde and N. Christophersen. Shrinkage Structure of Partial Least Squares. *Scandinavian Journal of Statistics*, 27:459–473, 2000.
22. N.J. Lobaugh, R. West, and A.R. McIntosh. Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares. *Psychophysiology*, 38:517–530, 2001.
23. R. Manne. Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration. *Chemometrics and Intelligent Laboratory Systems*, 2:187–197, 1987.
24. K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1997.
25. M. Martens and H. Martens. Partial Least Squares Regression. In J.R. Piggott, editor, *Statistical Procedures in Food Research*, pages 293–359. Elsevier Applied Science, London, 1986.
26. W.F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60:234–256, 1965.
27. M. Momma. Efficient Computations via Scalable Sparse Kernel Partial Least Squares and Boosted Latent Features. In *Proceedings of SIGKDD International Conference on Knowledge and Data Mining*, pages 654–659, Chicago, IL, 2005.
28. D.V. Nguyen and D.M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50, 2002.
29. J. Nilsson, S. de Jong, and A.K. Smilde. Multiway Calibration in 3D QSAR. *Journal of Chemometrics*, 11:511–524, 1997.
30. B. Parlett. *The symmetric eigenvalue problem*. SIAM, 1998.
31. A. Phatak and F. de Hoog. Exploiting the connection between PLS, Lanczos, and conjugate gradients: Alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16:361–367, 2003.
32. A. Phatak, P.M. Rilley, and A. Penlidis. The asymptotic variance of the univariate PLS estimator. *Linear Algebra and its Applications*, 354:245–253, 2002.
33. S. Rännar, F. Lindgren, P. Geladi, and S. Wold. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Chemometrics and Intelligent Laboratory Systems*, 8:111–125, 1994.
34. R. Rosipal and L.J. Trejo. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, 2:97–123, 2001.
35. R. Rosipal, L.J. Trejo, and B. Matthews. Kernel PLS-SVC for Linear and Nonlinear Classification. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 640–647, Washington, DC, 2003.

36. P.D. Sampson, A. P. Streissguth, H.M. Barr, and F.L. Bookstein. Neurobehavioral effects of prenatal alcohol: Part II. Partial Least Squares analysis. *Neurotoxicology and tetralogy*, 11:477–491, 1989.
37. C. Saunders, A. Gammerman, and V. Vovk. Ridge Regression Learning Algorithm in Dual Variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521, Madison, WI, 1998.
38. B. Schölkopf and A. J. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, 2002.
39. A. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley, 2004.
40. M. Stone and R.J. Brooks. Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society: B*, 52:237–269, 1990.
41. H. D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4:147–166, 1976.
42. G. Wahba. *Splines Models of Observational Data*, volume 59 of *Series in Applied Mathematics*. SIAM, Philadelphia, 1990.
43. J. Wallerius, L.J. Trejo, R. Matthew, R. Rosipal, and J.A. Caldwell. Robust feature extraction and classification of EEG spectra for real-time classification of cognitive state. In *Proceedings of 11th International Conference on Human Computer Interaction*, Las Vegas, NV, 2005.
44. L.E. Wangen and B.R. Kowalsky. A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, 3:3–20, 1989.
45. J.A. Wegelin. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical report, Department of Statistics, University of Washington, Seattle, 2000.
46. J. Westerhuis, T. Kourti, and J. MacGregor. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12:301–321, 1998.
47. H. Wold. Path models with latent variables: The NIPALS approach. In H.M. Blalock et al., editor, *Quantitative Sociology: International perspectives on mathematical and statistical model building*, pages 307–357. Academic Press, 1975.
48. H. Wold. Soft modeling: the basic design and some extensions. In J.-K. Jöreskog and H. Wold, editor, *Systems Under Indirect Observation*, volume 2, pages 1–53. North Holland, Amsterdam, 1982.
49. H. Wold. Partial least squares. In S. Kotz and N.L. Johnson, editors, *"Encyclopedia of the Statistical Sciences"*, volume 6, pages 581–591. John Wiley & Sons, 1985.
50. S. Wold. Nonlinear partial least squares modeling II, Spline inner relation. *Chemolab*, 14:71–84, 1992.
51. S. Wold, N. Kettaneh-Wold, and B. Skagerberg. Nonlinear PLS modelling. *Chemometrics and Intelligent Laboratory Systems*, 7:53–65, 1989.
52. S. Wold, H. Ruhe, H. Wold, and W.J. Dunn III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverse. *SIAM Journal of Scientific and Statistical Computations*, 5:735–743, 1984.
53. K.J. Worsley. An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5:254–258, 1997.