Kernel PLS-SVC for Linear and Nonlinear Classification

Roman Rosipal RROSIPAL@MAIL.ARC.NASA.GOV
NASA Ames Research Center, Computational Sciences Division, Moffett Field, CA 94035 USA
and
Department of Theoretical Methods, Slovak Academy of Sciences, Bratislava 842 19, Slovak Republic

Leonard J Trejo LTREJO@MAIL.ARC.NASA.GOV
NASA Ames Research Center, Computational Sciences Division, Moffett Field, CA 94035 USA

Bryan Matthews BMATTHEWS@MAIL.ARC.NASA.GOV
NASA Ames Research Center, Computational Sciences Division, Moffett Field, CA 94035 USA

Abstract

A new method for classification is proposed. This is based on kernel orthonormalized partial least squares (PLS) dimensionality reduction of the original data space followed by a support vector classifier. Unlike principal component analysis (PCA), which has previously served as a dimension reduction step for discrimination problems, orthonormalized PLS is closely related to Fisher's approach to linear discrimination or equivalently to canonical correlation analysis. For this reason orthonormalized PLS is preferable to PCA for discrimination. Good behavior of the proposed method is demonstrated on 13 different benchmark data sets and on the real world problem of classifying finger movement periods from non-movement periods based on electroencephalograms.

1. Introduction

The partial least squares (PLS) method (Wold, 1975; Wold et al., 1984) has been a popular modeling, regression, discrimination and classification technique in its domain of origin—chemometrics. In its general form, PLS creates score vectors (components, latent vectors) by using the existing correlations between different sets of variables (blocks of data) while also keeping most of the variance of both sets. PLS has proven to be useful in situations where the number of observed variables is much greater than the number of observations and high multicollinearity among the variables exists. This situation is also quite common in the case of kernel-based learning where the original data are mapped to a high-dimensional feature space corresponding to a reproducing kernel Hilbert space (RKHS). Motivated by the recent results in kernel-based learning and support vector machines (Vapnik, 1998; Schölkopf & Smola, 2002) a new method for classification is proposed. This is based on the kernel orthonormalized PLS method for dimensionality reduction combined with a support vector machine for classification (SVC) (Vapnik, 1998; Schölkopf & Smola, 2002).

Consider the ordinary least squares regression with outputs $\mathbf{Y}$ to be an indicator vector coding two classes with two different labels representing class membership. The regression coefficient vector from the least squares solution is then proportional to the linear discriminant analysis (LDA) direction (Hastie et al., 2001). This close connection between LDA and least square regression partially justified the use of PLS for discrimination. However, showing the close connection between Fisher's LDA, canonical correlation analysis (CCA) and orthonormalized PLS methods, Barker and Rayens (2003) more rigorously justified the use of PLS for discrimination. This connection also shows the preference of using orthonormalized PLS or its nonlinear kernel variant for dimensionality reduction in comparison to linear or nonlinear kernel-based principal components analysis (PCA) for discrimination.

In comparison to PLS regression on the dummy matrix $\mathbf{Y}$, the use of SVC on selected PLS score vectors is motivated by the possibility of constructing an *optimal separating hyperplane*, a better control for overlap between classes when the data are not separable, using a theoretically more principled "hinge" loss function instead of a squared-error loss function and finally to avoid the problem of *masking* of the classes in multiclass classification (Vapnik, 1998; Hastie et al., 2001).

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

Alternatively, other methods for classification (for example, LDA, logistic regression) applied on extracted PLS score vectors can be considered.

2. RHKS - basic definitions

A RKHS is uniquely defined by a positive definite kernel function $K(\mathbf{x},\mathbf{y})$; that is, a symmetric function of two variables satisfying the Mercer theorem conditions (Schölkopf & Smola, 2002). Consider $K(.,.)$ to be defined on a compact domain $\mathcal{X}\times\mathcal{X}$, $\mathcal{X}\subset R^N$. The fact that for any such positive definite kernel there exists a unique RKHS is well established by the *Moore-Aronszjan theorem*. The form $K(\mathbf{x},\mathbf{y})$ has the following *reproducing property*

$$f(\mathbf{x}) = \langle f(\mathbf{y}), K(\mathbf{x},\mathbf{y})\rangle_{\mathcal{H}} \quad \forall f\in\mathcal{H}$$

where $\langle .,.\rangle_{\mathcal{H}}$ is the scalar product in $\mathcal{H}$. The function K is called *a reproducing kernel* for $\mathcal{H}$.

It follows from Mercer's theorem that each positive definite kernel $K(\mathbf{x},\mathbf{y})$ can be written in the form

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{S}\lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{y}) \quad S\le\infty \tag{1}$$

where $\{\phi_i(.)\}_{i=1}^S$ are the eigenfunctions of the integral operator $\Gamma_K : L_2(\mathcal{X})\to L_2(\mathcal{X})$

$$(\Gamma_K f)(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{y})K(\mathbf{x},\mathbf{y})d\mathbf{y} \quad \forall f\in L_2(\mathcal{X})$$

and $\{\lambda_i>0\}_{i=1}^S$ are the corresponding positive eigenvalues. Rewriting (1) in the form

$$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{S}\sqrt{\lambda_i}\phi_i(\mathbf{x})\sqrt{\lambda_i}\phi_i(\mathbf{y}) = \Phi(\mathbf{x})^T\Phi(\mathbf{y}) \tag{2}$$

it becomes clear that any kernel $K(\mathbf{x},\mathbf{y})$ also corresponds to a canonical (Euclidean) dot product in a possibly high-dimensional space $\mathcal{F}$ where the input data are mapped by

$$\begin{aligned}\Phi:\ & \mathcal{X}\to\mathcal{F}\\ & \mathbf{x}\to(\sqrt{\lambda_1}\phi_1(\mathbf{x}),\sqrt{\lambda_2}\phi_2(\mathbf{x}),\ldots,\sqrt{\lambda_S}\phi_S(\mathbf{x}))\end{aligned}$$

The space $\mathcal{F}$ is usually denoted as a *feature space* and $\{\{\sqrt{\lambda_i}\phi_i(\mathbf{x})\}_{i=1}^S,\mathbf{x}\in\mathcal{X}\}$ as *feature mappings*. The number of basis functions $\phi_i(.)$ also defines the dimensionality of $\mathcal{F}$.

3. Partial Least Squares

Because the PLS technique is not widely known, first a description of linear PLS is provided which will simplify the description of its nonlinear kernel-based variant (Rosipal & Trejo, 2001).

3.1. Linear Partial Least Squares

Consider a general setting of the linear PLS algorithm to model the relation between two data sets (blocks of observed variables). Denote by $\mathbf{x}\in\mathcal{X}\subset\mathcal{R}^N$ an N-dimensional vector of variables in the first block of data and similarly $\mathbf{y}\in\mathcal{Y}\subset\mathcal{R}^M$ denotes a vector of variables from the second set. PLS models the relations between these two blocks by means of latent variables. 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

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{P}^T+\mathbf{F}\\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T+\mathbf{G}\end{aligned} \tag{3}$$

where the $\mathbf{T}$, $\mathbf{U}$ are $(n\times p)$ matrices of the extracted p 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 $\mathbf{F}$ and the $(n\times M)$ matrix $\mathbf{G}$ are the matrices of residuals. The PLS method, which in its classical form is based on the nonlinear iterative partial least squares (NIPALS) algorithm (Wold, 1975), finds weight vectors $\mathbf{w},\mathbf{c}$ such that

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

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

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

After the convergence, by regressing $\mathbf{X}$ on $\mathbf{t}$ and $\mathbf{Y}$ on $\mathbf{u}$, the loading vectors $\mathbf{p} = (\mathbf{t}^T\mathbf{t})^{-1}\mathbf{X}^T\mathbf{t}$ and $\mathbf{q} = (\mathbf{u}^T\mathbf{u})^{-1}\mathbf{Y}^T\mathbf{u}$ can be computed.

However, it can be shown that the weight vector $\mathbf{w}$ also corresponds to the first eigenvector of the following eigenvalue problem (Höskuldsson, 1988)

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \tag{4}$$

The X-scores $\mathbf{t}$ are then given as

$$\mathbf{t} = \mathbf{X}\mathbf{w} \tag{5}$$

Similarly, eigenvalue problems for the extraction of $\mathbf{t}$,$\mathbf{u}$ and $\mathbf{c}$ estimates can be derived (Höskuldsson, 1988). The nonlinear kernel PLS method is based on mapping

the original input data into a high-dimensional feature space $\mathcal{F}$. In this case the vectors $\mathbf{w}$ and $\mathbf{c}$ cannot be usually computed. Thus, the NIPALS algorithm needs to be reformulated into its kernel variant (Lewi, 1995; Rosipal & Trejo, 2001). Alternatively, the score vector $\mathbf{t}$ can be directly estimated as the first eigenvector of the following eigenvalue problem (Höskuldsson, 1988) (this can be easily shown by multiplying both sides of (4) by $\mathbf{X}$ matrix and using (5))

$$\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t} \tag{6}$$

The Y-scores $\mathbf{u}$ are then estimated as

$$\mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t} \tag{7}$$

3.2. Nonlinear Kernel Partial Least Squares

Now, consider a nonlinear transformation of $\mathbf{x}$ into a feature space $\mathcal{F}$. Using the straightforward connection between a RKHS and $\mathcal{F}$, Rosipal and Trejo (2001) have extended the linear PLS model into its nonlinear kernel form. Effectively this extension represents the construction of a linear PLS model in $\mathcal{F}$. Denote $\mathbf{\Phi}$ as the $(n \times S)$ matrix of mapped $\mathcal{X}$-space data $\Phi(\mathbf{x})$ into an S-dimensional feature space $\mathcal{F}$. Instead of an explicit mapping of the data, property (2) can be used resulting in

$$\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$$

where $\mathbf{K}$ represents the $(n \times n)$ *kernel Gram matrix* of the cross dot products between all input data points $\{\Phi(\mathbf{x}_i)\}_{i=1}^n$; that is, $\mathrm{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ where $K(.,.)$ is a selected kernel function. Similarly, consider a mapping of the second set of variables $\mathbf{y}$ into a feature space $\mathcal{F}_1$ and denote by $\mathbf{\Psi}$ the $(n \times S_1)$ matrix of mapped $\mathcal{Y}$-space data $\Psi(\mathbf{y})$ into an S_1-dimensional feature space $\mathcal{F}_1$. Analogous to $\mathbf{K}$ define the $(n \times n)$ kernel Gram matrix $\mathbf{K}_1$

$$\mathbf{K}_1 = \mathbf{\Psi}\mathbf{\Psi}^T$$

given by the kernel function $K_1(.,.)$. Using this notation the estimates of $\mathbf{t}$ (6) and $\mathbf{u}$ (7) can be reformulated into its nonlinear kernel variant

$$\begin{aligned} \mathbf{K}\mathbf{K}_1\mathbf{t} &= \lambda\mathbf{t} \\ \mathbf{u} &= \mathbf{K}_1\mathbf{t} \end{aligned} \tag{8}$$

Similar to linear PLS, a zero mean nonlinear kernel PLS model is assumed. To centralize the mapped data in a feature space $\mathcal{F}$ the following procedure must be applied (Schölkopf et al., 1998; Rosipal & Trejo, 2001)

$$\mathbf{K} \leftarrow (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{K}(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) \tag{9}$$

where $\mathbf{I}_n$ is an n-dimensional identity matrix and $\mathbf{1}_n$ represents a $(n \times 1)$ vector with elements equal to one. The same is true for $\mathbf{K}_1$.

After the extraction of new score vectors $\mathbf{t}, \mathbf{u}$ the matrices $\mathbf{K}$ and $\mathbf{K}_1$ are deflated by subtracting their rank-one approximations based on $\mathbf{t}$ and $\mathbf{u}$. The different forms of deflation correspond to different forms of PLS (see Wegelin (2000) for a review). Because (4) corresponds to the singular value decomposition of the transposed cross-product matrix $\mathbf{X}^T\mathbf{Y}$, computation of all eigenvectors from (4) at once involves a sequence of implicit rank-one deflations of the overall cross-product matrix. Although the weight vectors $\{\mathbf{w}_i\}_{i=1}^p$ will be mutually orthogonal the corresponding score vectors $\{\mathbf{t}_i\}_{i=1}^p$, in general, will not be mutually orthogonal. The same is true for the weight vectors $\{\mathbf{c}_i\}_{i=1}^p$ and the score vectors $\{\mathbf{u}_i\}_{i=1}^p$. This form of PLS was used by Sampson et al. (1989) and in accordance with Wegelin (2000) it is denoted as PLS-SB. The kernel analog of PLS-SB results from the computation of all eigenvectors of (8) at once. PLS1 (one of the blocks has single variable) and PLS2 (both blocks are multidimensional) generally used as regression methods use a different form of deflation. The deflation in the case of PLS1 and PLS2 is based on rank-one reduction of the $\mathbf{\Phi}$ and $\mathbf{\Psi}$ matrices using a new extracted score vector $\mathbf{t}$ at each step. It can be written in the kernel form for $\mathbf{K}$ matrix as follows (Rosipal & Trejo, 2001)

$$\mathbf{K} \leftarrow (\mathbf{I}_n - \mathbf{t}\mathbf{t}^T)\mathbf{K}(\mathbf{I}_n - \mathbf{t}\mathbf{t}^T)$$

and in the same way for $\mathbf{K}_1$. This deflation is based on the fact that the $\mathbf{\Phi}$ matrix is deflated as $\mathbf{\Phi} \leftarrow \mathbf{\Phi} - \mathbf{t}\mathbf{p}^T = \mathbf{\Phi} - \mathbf{t}\mathbf{t}^T\mathbf{\Phi}$, where $\mathbf{p}$ is the vector of loadings corresponding to the extracted unit norm score vector $\mathbf{t}$. Similarly for the $\mathbf{\Psi}$ matrix the deflation has the form $\mathbf{\Psi} \leftarrow \mathbf{\Psi} - \mathbf{t}\mathbf{c}^T = \mathbf{\Psi} - \mathbf{t}\mathbf{t}^T\mathbf{\Psi}$. In the case of PLS1 and PLS2 score vectors $\{\mathbf{t}_i\}_{i=1}^p$ are mutually orthogonal. In general, this is not true for $\{\mathbf{u}_i\}_{i=1}^p$ (Höskuldsson, 1988).

4. Fisher's LDA, CCA and PLS

Consider a set of N-dimensional 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 $\mathbf{0}_{n_i}$ is a $(n_i \times 1)$ vector of all zeros. Let $\mathbf{S}_x = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$, $\mathbf{S}_y = \frac{1}{n-1}\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{S}_{xy} = \frac{1}{n-1}\mathbf{X}^T\mathbf{Y}$ to be the sample estimates of $\mathcal{X}$ and $\mathcal{Y} \subset$

$\mathcal{R}^{g-1}$ space 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 $\mathbf{H} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$ and $\mathbf{E} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_i^j - \bar{\mathbf{x}}_i)(\mathbf{x}_i^j - \bar{\mathbf{x}}_i)^T$ represent the *among-classes* and *within-classes* sums-of-squares, where $\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_i^j$, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathbf{x}_i^j$ and $\mathbf{x}_i^j$ represents a N-dimensional vector for the j^{th} sample in the i^{th} class.

CCA is a method which finds a pair of linear transformations of each block of data with maximal correlation coefficient. This can be formally described as the maximization problem

$$\begin{aligned}&\max_{\mathbf{r}^T\mathbf{\Sigma}_x\mathbf{r}=\mathbf{s}^T\mathbf{\Sigma}_y\mathbf{s}=1}[corr(\mathbf{Xr},\mathbf{Ys})]^2 = \\ &= [corr(\mathbf{Xa},\mathbf{Yb})]^2 = \\ &= [cov(\mathbf{Xa},\mathbf{Yb})]^2/[var(\mathbf{Xa})var(\mathbf{Yb})]\end{aligned}$$

where similar to our previous notation the symbols *corr* and *var* denote the sample correlation and variance, respectively. An estimate of the weight vector $\mathbf{a}$ is given as the solution of the following eigenvalue problem (Mardia et al., 1997)

$$\mathbf{S}_x^{-1}\mathbf{S}_{xy}\mathbf{S}_y^{-1}\mathbf{S}_{yx}\mathbf{a} = \lambda\mathbf{a}$$

where the eigenvalues λ corresponds to the squared canonical correlation coefficient.

Without the assumption of Gaussian distribution of individual classes, Fisher developed a discrimination method based on a linear projection of the input data such that among-classes variance is maximized 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 two-class discrimination with multinormal distributions with the same covariance matrices, Fisher's LDA finds the same discrimination direction as LDA using Bayes theorem to estimate posterior class probabilities—the method providing the discrimination rule with minimal expected misclassification error (Mardia et al., 1997; Hastie et al., 2001).

The connection between Fisher's LDA directions and the directions given by CCA using a dummy matrix $\mathbf{Y}$ for group membership was first recognized by Bartlett. This connection expressed using the previously defined notation was formulated by Barker and Rayens (2003) (see also 11.5.4, Mardia et al., 1997) in the following two theorems:

Theorem 1

$$\mathbf{S}_{xy}\mathbf{S}_y^{-1}\mathbf{S}_{xy}^T = \tfrac{1}{n-1}\mathbf{H}$$

Theorem 2

$$\mathbf{S}_x^{-1}\mathbf{S}_{xy}\mathbf{S}_y^{-1}\mathbf{S}_{xy}^T\mathbf{a} = \lambda\mathbf{a} \Leftrightarrow \mathbf{E}^{-1}\mathbf{H}\mathbf{a} = \tfrac{\lambda}{1-\lambda}\mathbf{a}$$

The proof of the first theorem can by found in Barker and Rayens (2003). Using the property of the generalized eigenvalue problem, Theorem 1 and the fact that $(n-1)\mathbf{S}_x = \mathbf{E}+\mathbf{H}$, the second theorem can be proved.

A very close connection between Fisher's LDA, CCA and PLS methods for multi-class discrimination has been shown in Barker and Rayens (2003). This connection is based on the fact that PLS can be seen as a form of penalized CCA

$$\begin{aligned}&[cov(\mathbf{t},\mathbf{u})]^2 = [cov(\mathbf{Xw},\mathbf{Yc})]^2 = \\ &= var(\mathbf{Xw})[corr(\mathbf{Xw},\mathbf{Yc})]^2 var(\mathbf{Yc})\end{aligned}$$

with penalties given by PCA in $\mathcal{X}$- and $\mathcal{Y}$-spaces. Barker and Rayens (2003) suggested to remove the not meaningful $\mathcal{Y}$-space penalty $var(\mathbf{Yc})$ in the PLS discrimination scenario. This modification in fact represents a special case of the previously proposed orthonormalized PLS method (Worsley, 1997) using the indicator matrix $\mathbf{Y}$. In this case (4) is transformed into the eigenvalue problem

$$\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \tag{10}$$

Using Theorem 1 and the fact that $\mathbf{S}_{xy} = (n-1)\mathbf{X}^T\mathbf{Y}$ and $\mathbf{S}_y = (n-1)\mathbf{Y}^T\mathbf{Y}$ the eigenvectors of (10) are equivalent to the eigensolutions of

$$\mathbf{H}\mathbf{w} = \lambda\mathbf{w} \tag{11}$$

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 Fisher's LDA.

5. Kernel PLS-SVC for Classification

The connection between CCA, Fisher's LDA and PLS motivates the use of the orthonormalized PLS method for discrimination. The kernel variant[1] of this approach will transform (8) into the following equations

$$\begin{aligned}&\mathbf{K}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{t} = \mathbf{K}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\mathbf{t} = \lambda\mathbf{t} \\ &\mathbf{u} = \tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\mathbf{t}\end{aligned} \tag{12}$$

where $\tilde{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$ represents a matrix of uncorrelated and normalized original output variables. Interestingly, in the case of two-class discrimination the direction of the first kernel orthonormalized PLS

[1]Both linear and nonlinear PLS are considered. In the case of linear kernel a feature space $\mathcal{F}$ is equivalent to $\mathcal{X}$ and in this case the kernel variant of PLS will be preferable only if $N > n$.

score vector $\mathbf{t}$ is identical with the first score vector found by either the kernel PLS1 or the kernel PLS-SB method. This immediately follows from the fact that $\mathbf{Y}^T\mathbf{Y}$ is a number in this case. In this two-class scenario $\mathbf{K}\mathbf{Y}\mathbf{Y}^T$ is of a rank one matrix and kernel PLS-SB extracts only one score vector $\mathbf{t}$. In contrast, kernel orthonormalized PLS (or equivalently kernel PLS1) can extract additional score vectors, up to the rank of $\mathbf{K}$, each being similar to directions computed with CCA and Fisher's LDA on deflated feature space matrices. This provides more principled dimensionality reduction in comparison to PCA based on the criterion of maximum data variation in the $\mathcal{F}$-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 kernel orthonormalized 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 $\mathbf{t}$. For simplicity consider this deflation scheme in the original input space: $\mathbf{X}_1 = (\mathbf{I}_n - \mathbf{t}\mathbf{t}^T)\mathbf{X} = \mathbf{P_d}\mathbf{X}$, $\tilde{\mathbf{Y}}_1 = \mathbf{P_d}\tilde{\mathbf{Y}}$, where $\mathbf{P_d} = \mathbf{P_d}^T\mathbf{P_d}$ represents a projection matrix. Using these deflated matrices $\mathbf{X}_1$ and $\tilde{\mathbf{Y}}_1$ the eigenproblem (10) can be written in the form $\mathbf{X}_1^T\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\mathbf{X}_1\mathbf{w} = \lambda\mathbf{w}$. Thus, similar to the previous two-class discrimination the solution of this eigenproblem can be interpreted as the solution of (11) using the among-classes sum-of-squares matrix now computed on deflated input space matrix $\mathbf{X}_1$.

Kernel variants of CCA and Fisher's LDA have been proposed (Lai & Fyfe, 2000; Mika et al., 1999). Although the same relations among CCA, Fisher's LDA and PLS in a feature space $\mathcal{F}$ can be considered, kernel CCA and kernel Fisher DA suffer from a singularity problem in the case of higher dimensionality $S > n$. Both algorithms at some point need to invert singular matrices which is avoided by using the regularization concept of adding a small ridge (jitter) parameter on the diagonal of those matrices. The connection between a regularized form of CCA, PLS and orthonormalized PLS was developed in the context of canonical ridge analysis by Vinod (1976).

On several classification problems the use of kernel PCA for dimensionality reduction or de-noising followed by linear SVC computed on the reduced $\mathcal{F}$-space data representation has shown good results in comparison to nonlinear SVC using the original data representation (Schölkopf & Smola, 2002; Schölkopf et al., 1998). However, the theory of the previous section suggest to replace the kernel PCA data preprocessing step with a more principled kernel orthonormalized PLS approach. In comparison to kernel Fisher DA this may become more suitable in the situation of non-Gaussian class distribution in a feature space $\mathcal{F}$ where more than $g-1$ discrimination directions may better define an overall discrimination rule. The advantage of using linear SVC as the follow up step is motivated by the construction of an optimal separating hyperplane in the sense of maximizing of the distance to the closest point from either class (Vapnik, 1998; Schölkopf & Smola, 2002). Moreover, when the data are not separable the SVC approach provides a way to control the extent of this overlap. Thus, the kernel orthonormalized PLS is combined with the ν-SVC or the C-SVC (Schölkopf & Smola, 2002) classifier and this methodology is denoted as kernel PLS-SVC. A pseudo code of the method is provided in the Appendix.

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

6. Experiments

The usefulness of the kernel PLS-SVC method was tested on several benchmark data sets of two-class classification and on a real world problem of discriminating and classifying finger movements from periods of non-movement based on electroencephalograms (EEG).

6.1. Benchmark Data Sets

The data sets used in Rätsch et al. (2001) and Mika et al. (1999) were chosen. The data sets are freely available and can be downloaded from `http://www.first.gmd.de/~raetsch`. The data sets consist of 100 different training and testing partitions (except Splice and Image, consisting of 20 partitions). In all cases the Gaussian kernel was used. The unknown parameters (width of the Gaussian kernel, number of PLS score vectors, ν and C parameters for ν-SVC and C-SVC, respectively) were selected based on the minimum classification error using five-fold cross validation (CV) on the first five training sets.

The results are summarized in Table 1. Very good behavior of kernel PLS-SVC method can be observed. The null hypothesis about equal means using the C-SVC and kernel PLS-SVC methods was tested using a paired t-test (the individual test set classification errors for kernel Fisher DA are not available). The nonparametric sign and Wilcoxon matched-pairs signed-ranks tests were also used to test null hypotheses about the *direction* and *size* of the differences within pairs. The significance level for all tests was set to $\alpha = 0.05$. On five data sets (Banana, Diabetes, Ringnorm, Twonorm, Waveform) the null hypothesis about equal means was rejected. The one-sided alternative of both nonparametric tests indicated lower classification

errors of the kernel PLS-SVC approach in all five cases and also on B. Cancer. The paired t-test did not reject the null hypothesis about equal means on the Heart data set, but the one-sided alternative of the nonparametric tests indicate lower classification errors using C-SVC. The number of selected kernel PLS components determined by the CV approach was lower than 10 except for the Image data set where 27 score vectors were used. A relatively large improvement in terms of averaged classification error over kernel Fisher DA can be seen in this case. Interestingly, this superiority of kernel PLS-SVC over kernel Fisher DA was also observed in the case when only one PLS score vector was used (German, Ringnorm, Twonorm) but not on the Heart data set.

Table 1. Comparison of the mean and standard deviation test set classification errors between kernel Fisher DA (KFD) (Mika et al., 1999), C-SVC (Rätsch et al., 2001) and kernel PLS-SVC (asterisks indicate data sets where C-SVC was used in contrast to ν-SVC used on the remaining data sets). The method with minimum averaged classification error is highlighted in bold. The last row represents the mean of the values computed as the ratio between the averaged classification error of a method and the averaged classification error of the best method on a particular data set minus one.

Data Set	KFD	C-SVC	KPLS-SVC
Banana	10.8±0.5	11.5±0.5	**10.5±0.4**
B.Cancer	25.8±4.6	26.0±4.7	**25.1±4.5***
Diabetes	23.2±1.6	23.5±1.7	**23.0±1.7**
German	23.7±2.2	23.6±2.1	**23.5±2.0**
Heart	16.1±3.4	**16.0±3.3**	16.5±3.6
Image	4.76±0.58	**2.96±0.60**	3.03±0.61
Ringnorm	1.49±0.12	1.66±0.12	**1.43±0.10**
F.Solar	33.2±1.7	**32.4±1.8**	**32.4±1.8**
Splice	**10.5±0.6**	10.9±0.7	10.9±0.8
Thyroid	**4.20±2.07**	4.80±2.19	4.39±2.10
Titanic	23.2±2.06	**22.4±1.0**	**22.4±1.1***
Twonorm	2.61±0.15	2.96±0.23	**2.34±0.11**
Waveform	9.86±0.44	9.88±0.43	**9.58±0.36**
mean %	7.2±16.4	6.1±8.2	1.1±1.7

6.2. Finger Movement Detection

In an experiment designed to detect finger movements using EEG subjects performed a self-paced single finger tap about every five seconds (Trejo et al., 2003). In four runs the subject was instructed to alternate between the pinkie and index fingers on a single hand. Half of those runs were left and half were right hand only. In two runs the subject was then instructed to alternate between both hands keeping the same time separation between taps. Each run contained approximately 50 single taps. 62-channel EEG and 2-channel electrooculogram were recorded using a Neuroscan 64 channel EEG cap with two 32 channel syn-amps sampled at 1000 Hz. The electromyogram was also recorded using two electrodes placed on each wrist. The raw EEG was cut into one-second intervals with 300 ms before the motion and 700 ms after the beginning of the motion. The intervals were down sampled from 1000 to 128 data points using the Matlab routine `resample`. Both right and left hand intervals were labeled as motion and classified against periods of non-motion of equal length using the kernel PLS-SVC classifier and ν-SVC classifier alone. The experiment with the same subject was repeated two times with an interval of 56 days between the sessions. These days are denoted Day 1 and Day 2, respectively. Due to impedance problems with one of the electrodes (O_1) during the second day session only 61 channels of EEG were used. A total of 225 periods of movement and 579 periods of non-movement was extracted for Day 1 and 288 movement periods versus 657 non-movement periods for Day 2. The dimensionality of each period was 7808 (61 electrodes times 128 time points).

The accuracy to classify both finger movement and non-movement periods on data measured during Day 2 was based on the linear kernel PLS-SVC model. The model was trained on Day 1 data. The same was done using Day 2 data to predict Day 1. The parameters for the kernel PLS-SVC models were estimated using 10-fold CV on each day's data separately. First, the number of PLS score vectors was fixed and the ν parameter was estimated. In Fig. 1 the dependence of the correct classification rate on the number of selected PLS score vectors is depicted. The asterisks indicate the correct classification rate when the final number of the PLS score vectors was determined using the CV approach. The graphs indicate that a classification accuracy of over 90% can be achieved. Using a range of ν values for ν-SVC a maximum correct classification rate for the Day 1 to Day 2 scenario was 93.0% and 90.7% for the Day 2 to Day 1 scenario. The results with the nonlinear kernel PLS-SVC model using Gaussian kernel do not indicate improvement in comparison to its linear variant.

In the proposed discrimination scenario the individual PLS score vectors can be considered as different spatio-temporal processes with respect to differentiation between movement and non-movement periods. In the case of linear kernel PLS the corresponding weight vectors $\mathbf{w}$ (eq. (4)) can be computed and their values reflect important time points and spatial locations with respect to discrimination. These weight vectors were plotted as scalp topographical maps at

different times relative to finger movement (for example, Fig. 2). Based on the close visual inspection of these maps 16 EEG channels were selected out of all 61 channels. The selected electrodes were located predominantly over the right-hand side sensori-motor area (white areas in Fig. 2). Four electrodes from the occipital area (two on each side) were also selected (posterior black areas in Fig. 2). The results using this reduced number of electrodes are plotted in Fig. 1. In both cases the plots indicate comparable results to those achieved with the full EEG montage. However, in the second case when the Day 2 to Day 1 prediction scenario was used a reduced setting of the electrodes has a tendency of overfitting for a higher number of used components. To further justify this electrode reduction 100 different training and testing partitions with the ratio of splits 40:60% were created. This was done for each day independently. Using 10-fold CV on both the full and reduced montage training partitions, linear kernel PLS-SVC models were compared in terms of correct classification rates achieved on 100 test partitions. For Day 1 the averaged correct classification for reduced set of electrodes was 89.6% $\pm$1.5 in comparison to 89.3% $\pm$1.4 using the full EEG montage. For Day 2 the results were 91.9% $\pm$1.3 and 92.2% $\pm$1.1, respectively. In both cases, using the paired t-test, the null hypothesis about equal means was not rejected (p-values > 0.05).

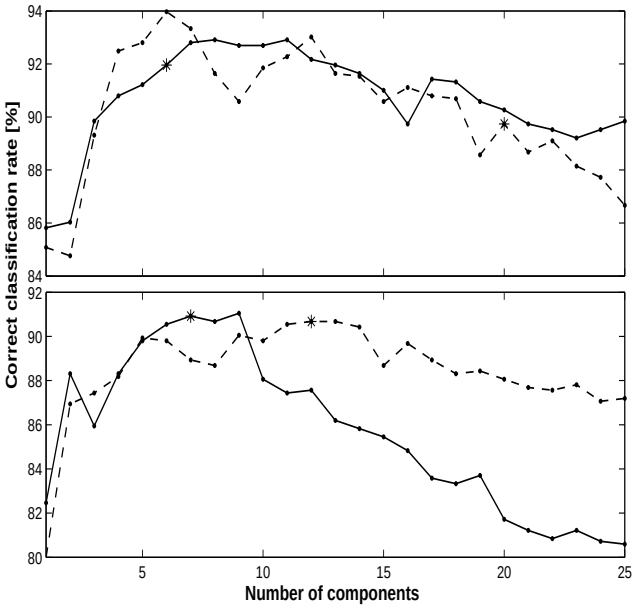

Figure 1. Comparison of the full 61 EEG electrodes setting (dashed line) with the reduced setting of 16 electrodes (solid line). Asterisks show results achieved in the case where 10-fold CV was used to select a number of score vectors and the ν parameter for ν-SVC. *Top:* the Day 1 to Day 2 scenario. *Bottom:* the Day 2 to Day 1 scenario.

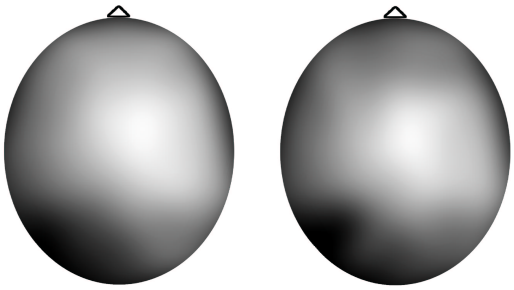

Figure 2. Topographic scalp projections of the first PLS weight vector at the time point 370ms after the onset of movement. *Left*: Day 1 data. *Right:* Day 2 data. (Top of the graph is the front of the head.)

7. Conclusions

A new kernel PLS-SVC classification technique was proposed. Results achieved on 13 benchmark data sets demonstrate usefulness of the proposed method and its competitiveness with other state-of-the-art classification methods. On six benchmark data sets a statistically significant superiority of kernel PLS-SVC over C-SVC was observed. In contrast, this tendency was observed only in one case for C-SVC. In terms of averaged classification error the superiority of kernel PLS-SVC over kernel Fisher DA was observed in 10 out of 13 benchmark data sets. In seven cases this was achieved using more than one score vector, which suggests, that a single direction extracted by kernel Fisher DA on these data sets is not adequate to discriminate two different classes.

In the case of finger movement detection from EEG, a linear kernel PLS approach provided a way to—in practice desirable—reduce the number of used electrodes without the degradation of the classification accuracy. It is the topic of a current more detailed study to analyze the individual spatio-temporal processes as defined by the extracted PLS score vectors. This would provide a more principled way for the selection of important spatial and temporal changes during the finger motion. The topographical maps constructed using the weight vectors of the constructed ν-SVC models (one weight vector for each model) have shown similarity between the plots using the weight vectors corresponding to the first PLS score vectors. However, these weight vectors represent a "global" discrimination of spatio-temporal processes. Moreover, they are computed using the support vectors only.

A theoretical connection between Fisher's LDA, CCA and PLS was described. This connection indicates that in the case of dimensionality reduction with respect to discrimination in $\mathcal{F}$ the kernel orthonormalized PLS method should be preferred over kernel PCA.

References

Barker, M., & Rayens, W. S. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*, 166–173.

de Jong, S., Wise, B. M., & Ricker, N. L. (2001). Canonical partial least squares and continuum power regression. *Journal of Chemometrics*, *15*, 85–100.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Höskuldsson, A. (1988). PLS Regression Methods. *Journal of Chemometrics*, *2*, 211–228.

Lai, P. L., & Fyfe, C. (2000). Kernel and Nonlinear Canonical Correlation Analysis. *International Journal of Neural Systems*, *10*, 365.

Lewi, P. J. (1995). Pattern recognition, reflection from a chemometric point of view. *Chemometrics and Intelligent Laboratory Systems*, *28*, 23–33.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1997). *Multivariate Analysis*. Academic Press.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K. R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX* (pp. 41–48).

Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. *Machine Learning*, *42*, 287–320.

Rosipal, R., & Trejo, L. J. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, *2*, 97–123.

Sampson, P. D., Streissguth, A. P., Barr, H. M., & Bookstein, F. L. (1989). Neurobehavioral effects of prenatal alcohol: Part II. Partial Least Squares analysis. *Neurotoxicology and tetralogy*, *11*, 477–491.

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press.

Schölkopf, B., Smola, A. J., & Müller, K. R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, *10*, 1299–1319.

Trejo, L. J., Wheeler, K., Jorgensen, C., Rosipal, R., Clanton, S., Matthews, B., Hibbs, A., Matthews, R., & Krupka, M. (in press, 2003). Multimodal Neuroelectric Interface Development. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

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

Wegelin, J. A. (2000). *A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case* (Technical Report). Department of Statistics, University of Washington, Seattle.

Wold, H. (1975). Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. In J. Gani (Ed.), *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, 520–540. Academic Press, London.

Wold, S., Ruhe, H., Wold, H., & Dunn III, W. J. (1984). 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.

Worsley, K. J. (1997). An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, *5*, 254–258.

Appendix

In the case of one-dimensional output SIMPLS algorithm provides the same solution than PLS1. Thus, for two-class classification a computationally more efficient SIMPLS algorithm can be used (de Jong et al., 2001). This is based on the fact that in this case $\mathbf{t} \propto \mathbf{KY}$. The kernel PLS-SVC algorithm can be then defined in three major steps:

1) kernel PLS components extraction
   compute $\mathbf{K}$ – centralized Gram matrix (9)
   set $\mathbf{K}_{res} = \mathbf{K}$, p - the number of score vectors
   for $i = 1$ to p
      $\mathbf{t}_i = \mathbf{K}_{res}\mathbf{Y}$
      $\|\mathbf{t}_i\| \rightarrow 1$
      $\mathbf{u}_i = \mathbf{Y}(\mathbf{Y}^T\mathbf{t}_i)$
      $\mathbf{K}_{res} \leftarrow \mathbf{K}_{res} - \mathbf{t}_i(\mathbf{t}_i^T\mathbf{K}_{res})$
      $\mathbf{Y} \quad \leftarrow \mathbf{Y} - \mathbf{t}_i(\mathbf{t}_i^T\mathbf{Y})$
   end
   $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p]$; $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p]$
2) projection of test samples (Rosipal & Trejo, 2001)
   compute $\mathbf{K}_t$ – centralized test set Gram matrix
   $\mathbf{T}_t = \mathbf{K}_t\mathbf{U}(\mathbf{T}^T\mathbf{K}\mathbf{U})^{-1}$
3) ν-SVC or C-SVC build on score vectors $\mathbf{T}$, $\mathbf{T}_t$

In the case of multi-class classification ($g > 2$) a kernel variant of the NIPALS algorithm (Rosipal & Trejo, 2001) with uncorrelated outputs $\tilde{\mathbf{Y}}$ or eigenproblem (12) has to be solved to extract $\{\mathbf{t}_i\}_{i=1}^p$ and $\{\mathbf{u}_i\}_{i=1}^p$.