Kernel PLS-SVC for Linear and Nonlinear Classification

Roman Rosipal[1,3], Leonard J. Trejo[1], Bryan Matthews[2]

[1]*NASA Ames Research Center,*
Computational Sciences Division, Moffett Field, CA
[2]*QSS Group Inc., NASA Ames Research Center,*
Computational Sciences Division, Moffett Field, CA
[3]*Department of Theoretical Methods*
Slovak Academy of Sciences, Bratislava, Slovak Republic

Outline

1. Introduction to PLS
2. PLS and Kernel PLS Discrimination
   (relation to CCA and Fisher's LDA)
3. Experimental Results

Partial Least Squares

- PLS - a class of techniques for modeling relations between blocks of observed variables by means of latent variables
- Herman Wold ('66,'81) - NIPALS - to linearize models nonlinear in the parameters
- Svante Wold et. al ('83) - extended PLS for the overdetermined regression problems
- Chemometrics - strong latent variable structure

Partial Least Squares

- data sets:

$$\mathbf{X}\ (n_{objects} \times N_{variables})$$
$$\mathbf{Y}\ (n_{objects} \times M_{responses})$$

  – *zero-mean*

- bilinear decomposition:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

where:

$\mathbf{T}, \mathbf{U}$ matrix of score variables (LV, components)

$\mathbf{P}, \mathbf{Q}$ matrix of loadings

$\mathbf{E}, \mathbf{F}$ matrix of residuals (errors)

- PLS - bilinear decomposition of $\mathbf{X}$ and $\mathbf{Y}$ maximizing covariance between score vectors $\mathbf{t} = \mathbf{Xw}$ and $\mathbf{u} = \mathbf{Yc}$

$$
\begin{aligned}
\max_{|\mathbf{r}|=|\mathbf{s}|=1}[cov(\mathbf{Xr}, \mathbf{Ys})]^2 &= [cov(\mathbf{Xw}, \mathbf{Yc})]^2 \\
&= var(\mathbf{Xw})[corr(\mathbf{Xw}, \mathbf{Yc})]^2 var(\mathbf{Yc}) \\
&= [cov(\mathbf{t}, \mathbf{u})]^2
\end{aligned}
$$

- $\boxed{\begin{array}{l} \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \\ \mathbf{t} = \mathbf{Xw} \end{array}}$ $\boxed{\begin{array}{l} \mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t} \\ \mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t} \end{array}}$
- $\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$; $\mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$
- deflation schemes define different forms of PLS

Forms of Partial Least Squares

- PLS1, PLS2: rank-one approximation on $\mathbf{X}, \mathbf{Y}$ with $\mathbf{t}$

  $\mathbf{X} \rightarrow \mathbf{X} - \mathbf{t}\mathbf{p}^T$; $\mathbf{Y} \rightarrow \mathbf{Y} - \mathbf{t}\mathbf{c}^T$

  mutually orthogonal components $\mathbf{t}_i$, $i = 1, \ldots, p$

- PLS-SB: SVD on $\mathbf{Y}^T\mathbf{X}$

  mutually orthogonal weight vectors $\mathbf{w}_i, \mathbf{c}_i$

  generally not orthogonal $\mathbf{t}_i$ and $\mathbf{u}_i$

- 1st SV$_{i+1} \geq$ 2nd SV$_i \rightarrow$ we select one component at a time

Partial Least Squares Discrimination

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

- orthonormalized PLS

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

$$\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} = \mathbf{I}$$

Orthonormalized PLS vs. CCA, Fisher's LDA

[Barker & Rayens 2003]

- orthonormalized PLS

  $$\max_{|\mathbf{r}|=|\mathbf{s}|=1}[cov(\mathbf{Xr}, \tilde{\mathbf{Y}}\mathbf{s})]^2 = var(\mathbf{Xw})[corr(\mathbf{Xw}, \tilde{\mathbf{Y}}\mathbf{c})]^2$$

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

  $$\boxed{\begin{array}{l}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{Xw} = \lambda\mathbf{w} \\ \mathbf{Hw} = \lambda\mathbf{w}\end{array}}$$

- CCA, Fisher's LDA

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

  $$\boxed{\begin{array}{l}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{Xa} = \lambda\mathbf{a} \\ \mathbf{E}^{-1}\mathbf{Ha} = \frac{\lambda}{1-\lambda}\mathbf{a}\end{array}}$$

Kernel PLS Discrimination

- linear PLS discrimination in a feature space $\mathcal{F}$
- nonlinear kernel-based PLS:

$$\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t} \qquad \Rightarrow \qquad \boxed{\mathbf{K}\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t}}$$
$$\mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t} \qquad \qquad \boxed{\mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t}}$$

- nonlinear kernel-based orthonormalized PLS:

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

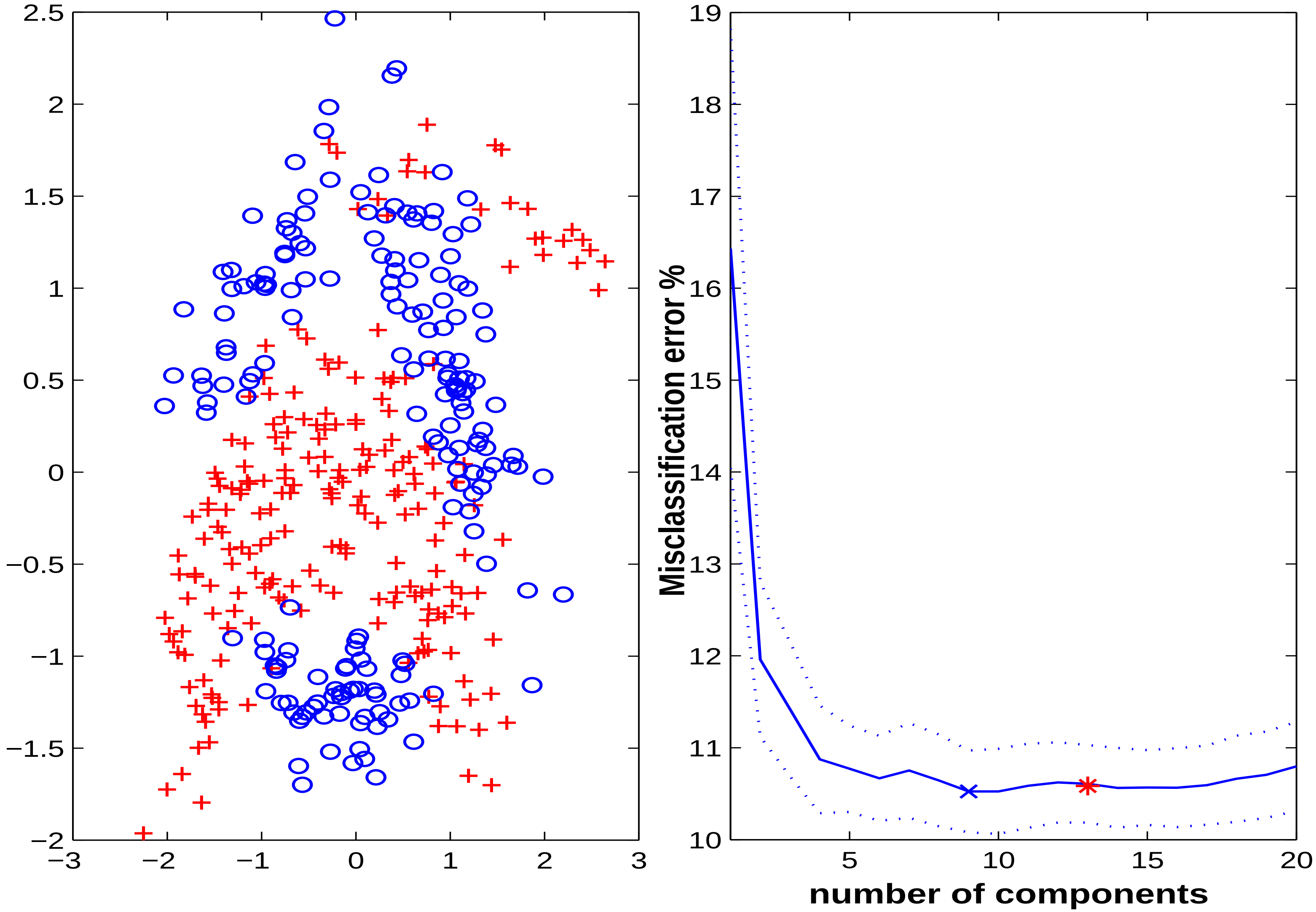
Kernel PLS-SVC Classification

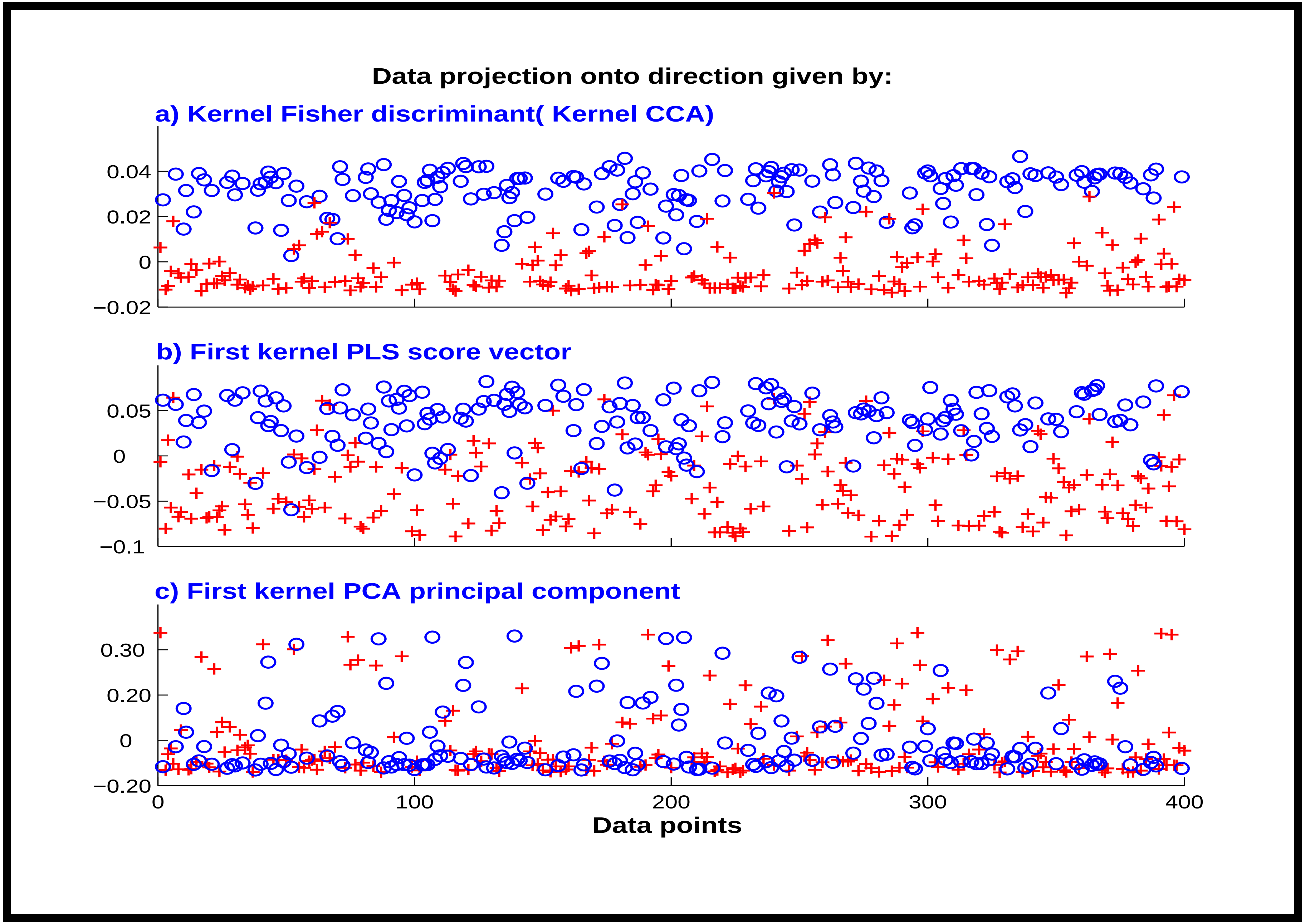
- orthonormalized kernel PLS + SVC (KPLS-SVC)
- orthonormalized kernel PLS can be combined with other existing classifiers (e.g. LDA, logistic regression)

Experiments:

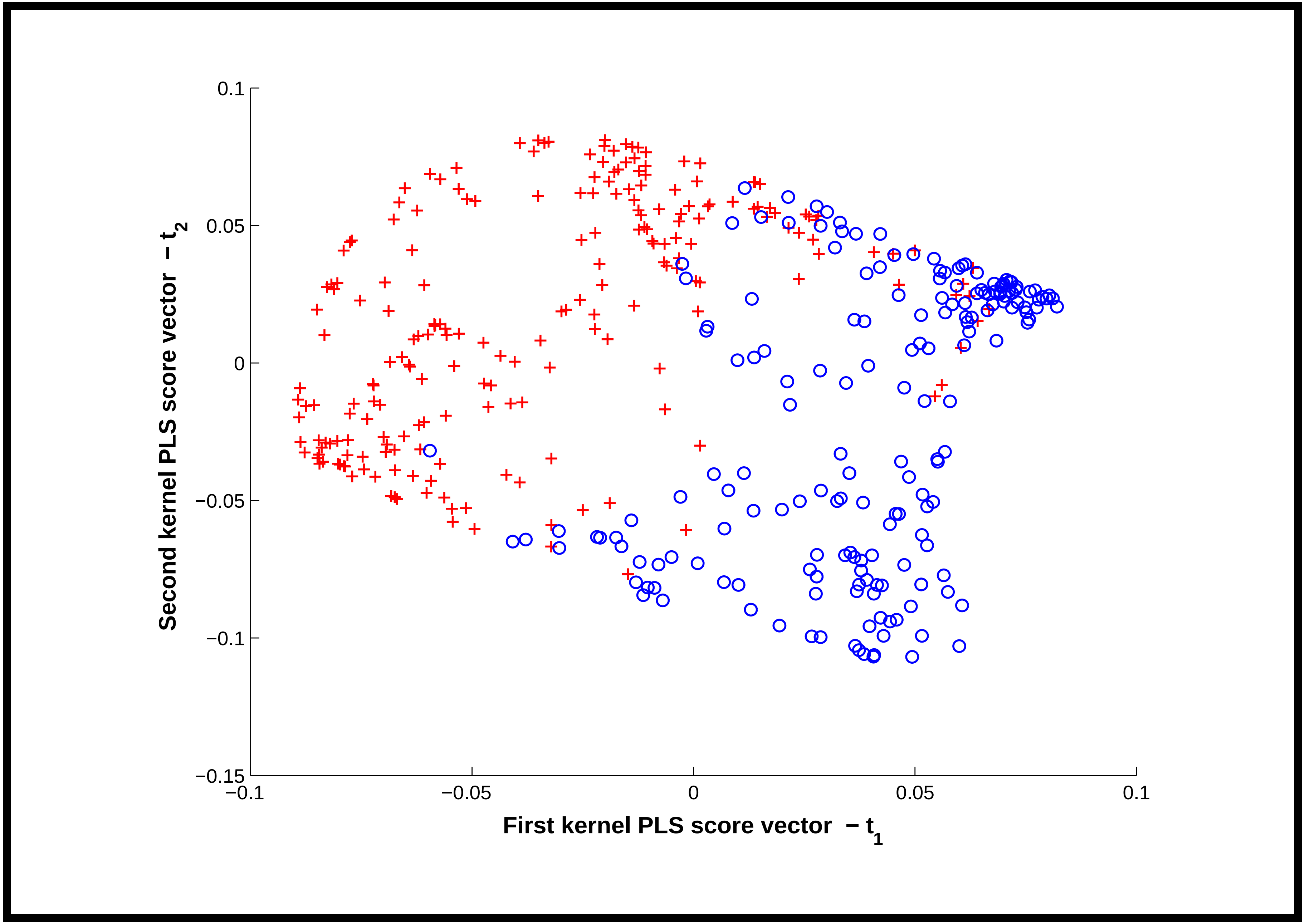
- 13 benchmark data sets of two-class classification problem
  `http://www.first.gmd.de/~raetsch`
- vowel sounds data set - multi-class problem (11 classes)
- classification of finger movement periods from non-movement periods based on electroencephalograms (EEG)

Banana data set

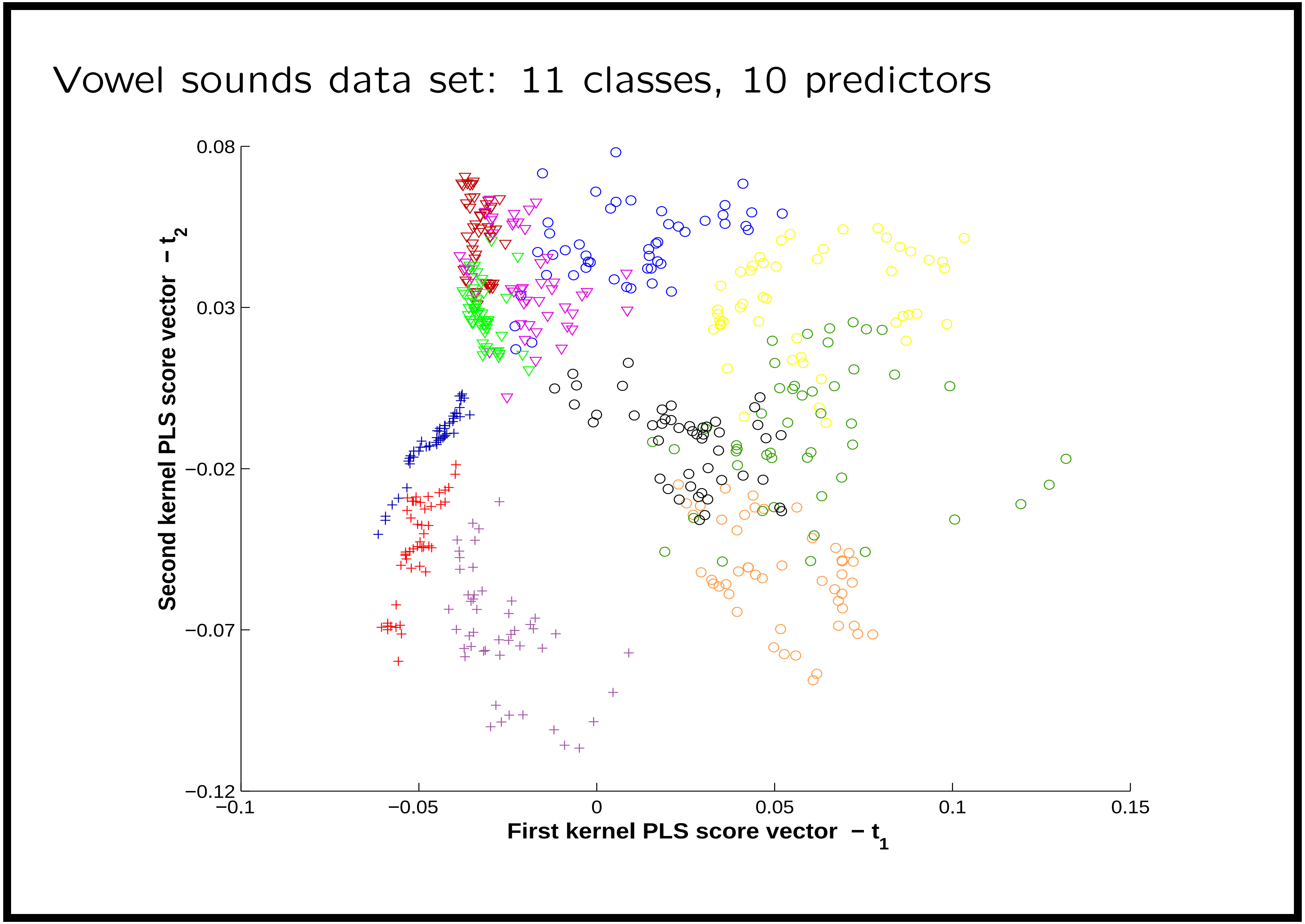




Data projection onto direction given by:
a) Kernel Fisher discriminant(Kernel CCA)
0.04
0.02
0
−0.02
b) First kernel PLS score vector
0.05
0
−0.05
−0.1
c) First kernel PCA principal component
0.30
0.20
0
−0.20
0
100
200
300
400
Data points


0.1
0.05
0
−0.05
−0.1
−0.15
−0.1
−0.05
0
0.05
0.1
Second kernel PLS score vector − t_2
First kernel PLS score vector − t_1

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±1.6**
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**

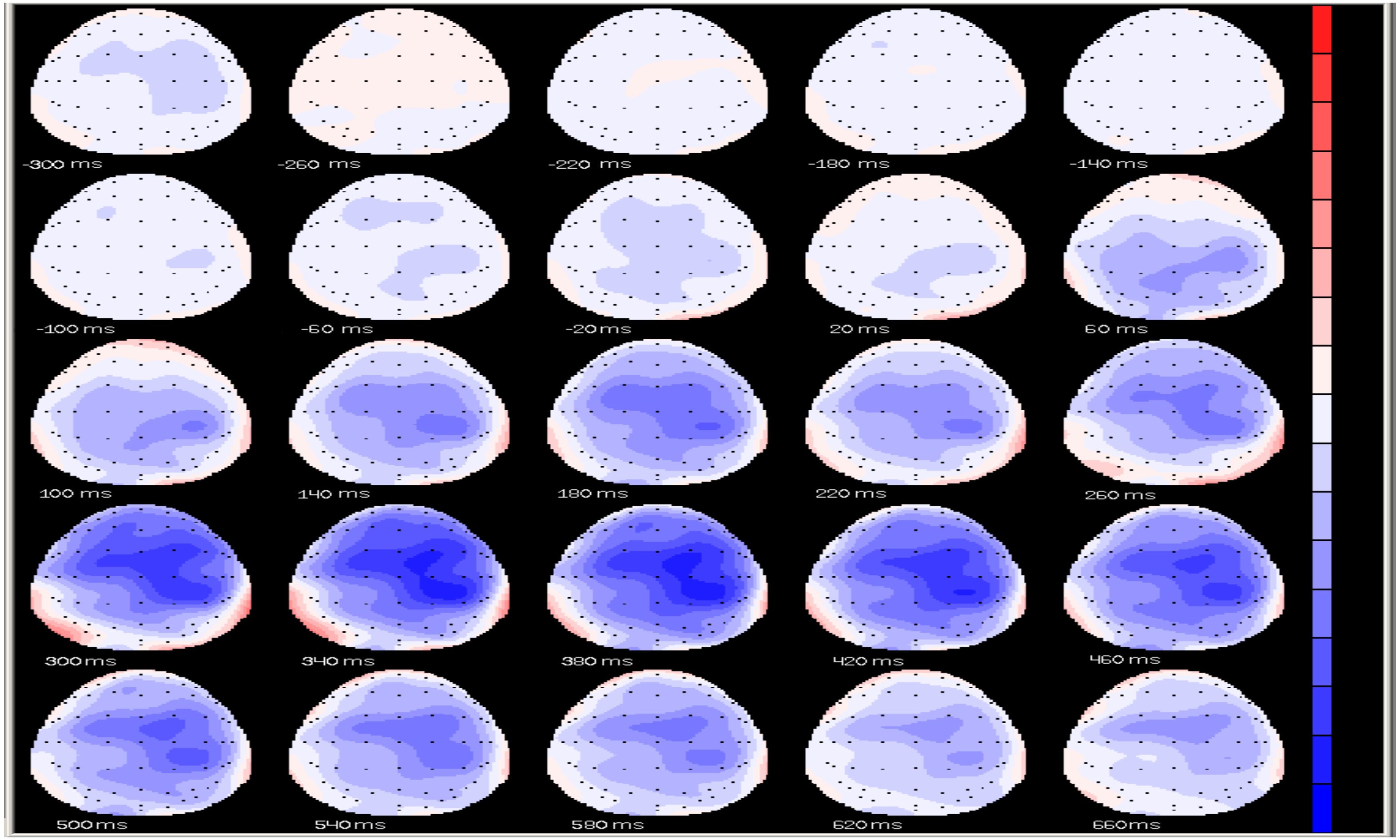


Vowel sounds data set: 11 classes, 10 predictors
0.08
0.03
−0.02
−0.07
−0.12
−0.1
−0.05
0
0.05
0.1
0.15
First kernel PLS score vector − t1
Second kernel PLS score vector − t2

Method	Training Error	Testing Error
LDA	0.32	0.56
SVC (linear) - 1vs1	0.19	0.51
KPLS-SVC (linear) - 1vs1	0.16	0.47
FDA/MARS (df=2)	0.02	0.42
FDA/MARS (df=6,red. dim.)	0.13	0.39
SVC (gauss) - 1vs1	0.01	0.37
KPLS-SVC (gauss) - 1vs1	0.01	0.35
SVC (gauss, w $\leq$ 5) - 1vs1	0.002	0.29
KPLS-SVC (gauss, w $\leq$ 5) - 1vs1	0.002	0.33

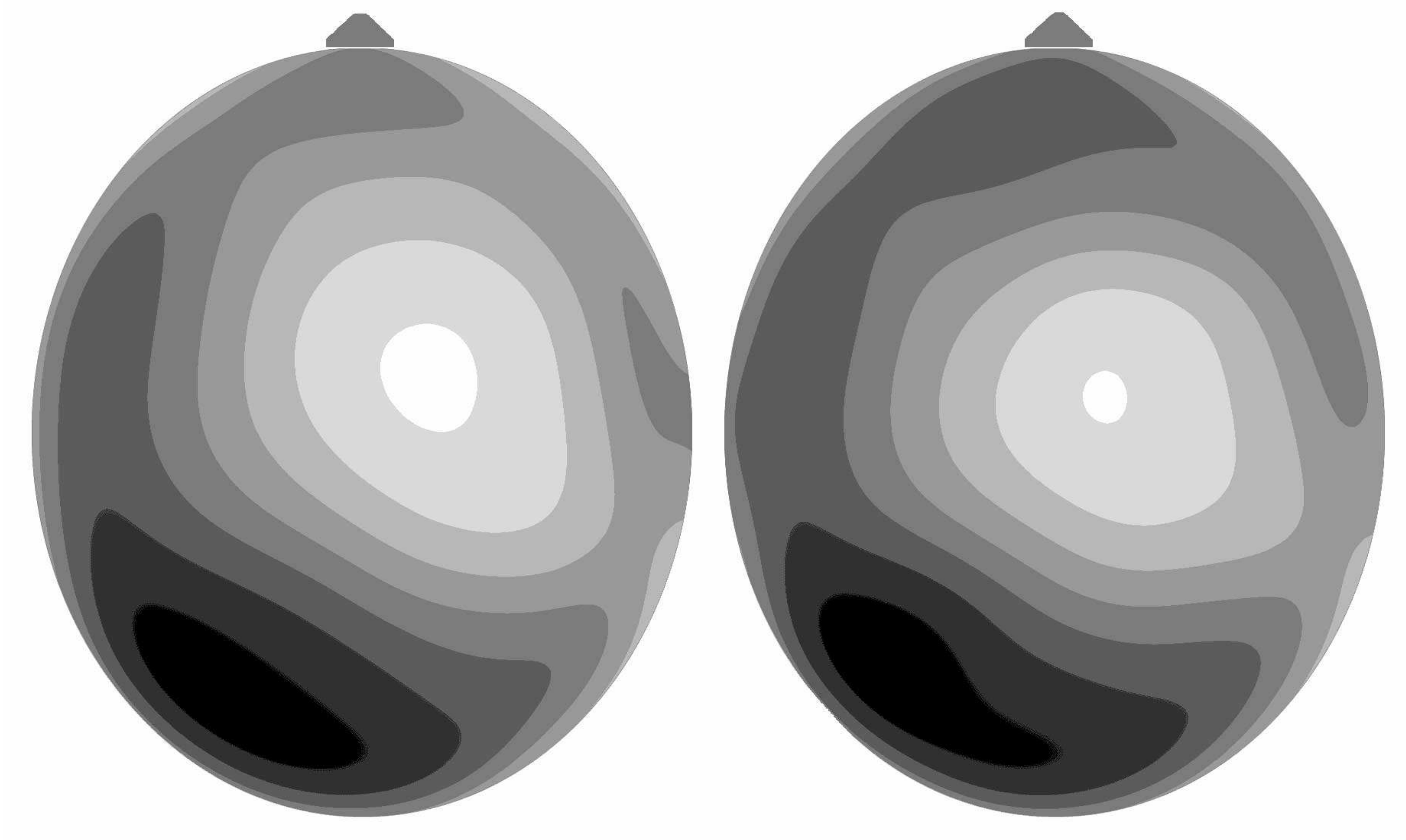
Finger movement periods vs. non-movement periods

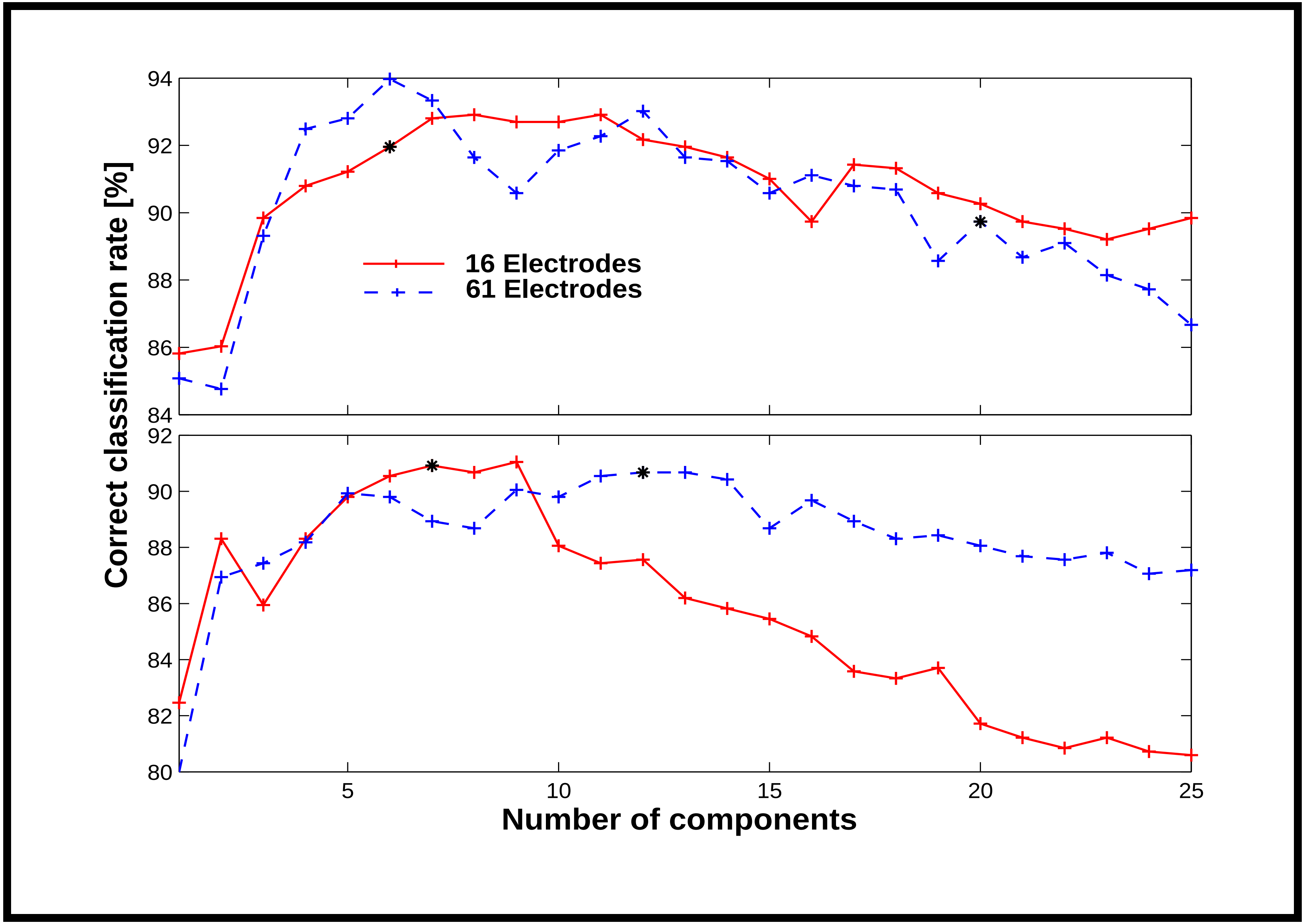
PLS-derived Spatio-temporal Filter - 01/09/2003



PLS-derived Spatio-temporal Filter
(370ms after button press)

11/14/2002 01/09/2003





Correct classification rate [%]
Number of components
16 Electrodes
61 Electrodes
94
92
90
88
86
84
92
90
88
86
84
82
80
5
10
15
20
25

Conclusions

- PLS discrimination - useful method for dimensionality reduction, visualization
- PLS discrimination preferred over PCA
- nonlinear kernel-based version of PLS discrimination provided
- KPLS-SVC achieved good results on the used data sets

?