Overview and Some Aspects of Partial Least Squares

Roman Rosipal

Austrian Research Institute for Artificial Intelligence
Vienna, Austria

(in collaboration with Leonard J. Trejo from
NASA Ames Research Center, CA)

Outline

1. History of PLS
2. Review of PLS and its Modifications
3. PLS Regression
4. "The Peculiar Shrinkage Properties" of PLS Regression
5. PLS for Discrimination/Classification
6. Experimental Results

History of Partial Least Squares

- PLS - a class of techniques for modeling relations between blocks of observed variables by means of latent variables
- Herman Wold'66,'75 - NIPALS - to linearize models nonlinear in the parameters
- Svante Wold et. al '83 - NIPALS extended for the overdetermined regression problems - PLS Regression
- Chemometrics - strong latent variable structure
- Math. Statistics - Stone & Brooks'90, Frank & Friedman'93, Garthwaite'94, Breiman & Friedman'97, etc.

- fMRI data
  - McIntosh et. al '96, Worsley'97, Nielsen et. al '98
- EEG, ERP data
  - Lobaugh et.al '01
  - Rosipal & Trejo'01 - nonlinear kernel PLS
- other applications
  - classification of microarray gene expression profiles (Nguyen & Rocke'02)
  - drug design (Bennett et. al '02,'03)
  - music data (Saunders et. al '04)

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 vectors (LV, components)

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

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

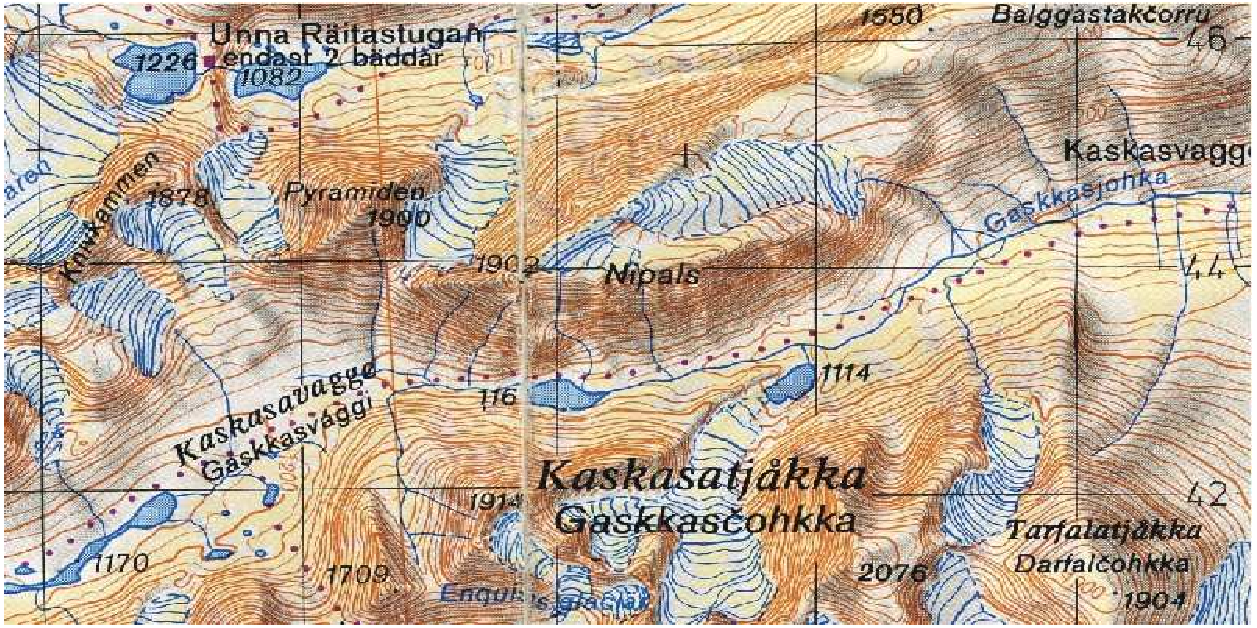
- PLS - bilinear decomposition of $\mathbf{X}$ and $\mathbf{Y}$ maximizing covariance between score vectors $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{c}$

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

- NIPALS algorithm finds the weights $\mathbf{w}, \mathbf{c}$:

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) $\mathbf{u} = \mathbf{Y}\mathbf{c}$
	7) *go to* 1)

- $\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$; $\mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$


1550
Balggastakčorru
46
Unna Räitastugan
endast 2 bäddar
1226
1082
Kaskasvagg
Gaskkasjohka
1878
Pyramiden
1900
Nipals
44
1114
116
Kaskasavagge
Gaskkasvággi
Kaskasatjåkka
Gaskkasčohkka
42
Tarfalatjåkka
Darfalčohkka
2076
1904
1709
1170

- instead of NIPALS we can solve an eigenproblem:

$$\mathbf{w} \propto \mathbf{X}^T\mathbf{u} \propto \mathbf{X}^T\mathbf{Y}\mathbf{c} \propto \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} \propto \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}$$

$$\boxed{\begin{array}{l}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \\ \mathbf{t} = \mathbf{X}\mathbf{w}\end{array}} \quad \text{or} \quad \boxed{\begin{array}{l}\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t} \\ \mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t}\end{array}}$$

- sequential extraction of $\{\mathbf{t}_i\}_{i=1}^m$
  $\mathbf{X}_0 = \mathbf{X}$
  $\mathbf{t}_i = \mathbf{X}_{i-1}\mathbf{w}_i \;\;,\;\; \mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{t}_i\mathbf{p}_i^T = \mathbf{X} - \sum_{j=1}^{i}\mathbf{t}_j\mathbf{p}_j^T$

- deflation schemes define different forms of PLS

Forms of Partial Least Squares

- PLS1, PLS2: rank-one approximation of $\mathbf{X}, \mathbf{Y}$ with a score vector $\mathbf{t}$ and vector of loadings $\mathbf{p}, \mathbf{q}$
  - $\mathbf{X} \to \mathbf{X} - \mathbf{t}\mathbf{p}^T$; $\mathbf{Y} \to \mathbf{Y} - \mathbf{t}\mathbf{c}^T$
  - mutually orthogonal score vectors $\mathbf{t}_i$, $i = 1, \ldots, m$
  - 1st SV$_{i+1} \geq$ 2nd SV$_i \to$ select one score vector at a time
- PLS Mode A: rank-one approximation of $\mathbf{X}, \mathbf{Y}$ with score vectors $\mathbf{t}, \mathbf{u}$ and vector of loadings $\mathbf{p}, \mathbf{q}$
  - $\mathbf{X} \to \mathbf{X} - \mathbf{t}\mathbf{p}^T$; $\mathbf{Y} \to \mathbf{Y} - \mathbf{u}\mathbf{q}^T$
  - mutually orthogonal score vectors $\mathbf{t}_i, \mathbf{u}_i$, $i = 1, \ldots, m$

- PLS-SB: SVD of $\mathbf{Y}^T\mathbf{X} = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^T$
  - $\mathbf{Y}^T\mathbf{X} \rightarrow \mathbf{Y}^T\mathbf{X} - \sigma\mathbf{a}\mathbf{b}^T$
  - mutually orthogonal weight vectors $\mathbf{a}_i, \mathbf{b}_i$
  - generally not orthogonal score vectors $\mathbf{c}_i = \mathbf{X}\mathbf{a}_i$, $\mathbf{d}_i = \mathbf{Y}\mathbf{b}_i$
- SIMPLS :(de Jong'93)
  - avoids deflation of $\mathbf{X}$; i.e. finds weight vectors $\tilde{\mathbf{w}}_i$ such that $\tilde{\mathbf{T}} = \mathbf{X}_0\tilde{\mathbf{W}}$
  - SVD of $\mathbf{X}_0^T\mathbf{Y}_0$ + constraint of mutually orthogonal $\tilde{\mathbf{t}}_i$
  - sequence of SVD problems $\tilde{\mathbf{P}}_i^{\perp}\mathbf{X}_0^T\mathbf{Y}_0$
    $\tilde{\mathbf{P}}_i^{\perp}$ an orthogonal projector onto $\tilde{\mathbf{P}}_i = [\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_i]$
    where $\tilde{\mathbf{p}}_i = \mathbf{X}_0^T\tilde{\mathbf{t}}_i/(\tilde{\mathbf{t}}_i^T\tilde{\mathbf{t}}_i)$ are loadings vectors
  - same as PLS1 but differs for PLS2
- Hinkel & Rayens'98-00; Frank & Friedman'93:
  - constraint maximization of covariance

CCA, PLS, and PCA $\rightleftharpoons$ CR

- PLS:

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

- CCA:

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

- PCA:

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

Canonical Ridge Analysis - CCA $\rightleftharpoons$ PLS

$$([1-\gamma_X]\mathbf{X}^T\mathbf{X}+\gamma_X\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}([1-\gamma_Y]\mathbf{Y}^T\mathbf{Y}+\gamma_Y\mathbf{I})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{w}=\lambda\mathbf{w}$$

- CCA: $\gamma_X = 0,\ \gamma_Y = 0$
- PLS: $\gamma_X = 1,\ \gamma_Y = 1$
- Orthonormalized PLS: $\gamma_X = 1,\ \gamma_Y = 0$ or $\gamma_X = 0,\ \gamma_Y = 1$
- Ridge Regression, Regularized FDA or CCA:
  $\gamma_X \in (0,1),\ \mathbf{Y} \in \mathcal{R}$

PLS Regression (PLS1, PLS2)

- assume: (i) $\mathbf{T}$ are good predictors of $\mathbf{Y}$

  (ii) the *inner loop* relation $\mathbf{U} = \mathbf{T} + \mathbf{H}$; i.e.

  $\mathbf{Y}$ is a linear function of $\mathbf{T}$

  H matrix of residuals (errors)

- linear PLS regression model:

  $\mathbf{Y} = \mathbf{T}\mathbf{C}^T + \mathbf{F}^* = \mathbf{X}\mathbf{B} + \mathbf{F}^*$, $\mathbf{F}^*$ matrix of residuals (errors)

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

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

- $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{C}^T = \mathbf{X}\mathbf{B}$
- using the existing relations among $\mathbf{t}, \mathbf{u}, \mathbf{c}, \mathbf{w}$:
  $\mathbf{B} = \mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y}$
- train data:
  $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} = \mathbf{T}\mathbf{T}^T\mathbf{Y} = \mathbf{T}\mathbf{C}^T$
  single output: $\hat{y}(\mathbf{x}) = c_1 t_1(\mathbf{x}) + c_2 t_2(\mathbf{x}) + \ldots + c_m t_m(\mathbf{x})$
- test data:
  $\hat{\mathbf{Y}}_t = \mathbf{X}_t\mathbf{X}^T\mathbf{U}(\mathbf{T}^T\mathbf{X}\mathbf{X}^T\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} = \mathbf{T}_t\mathbf{C}^T$

PLS1 $\rightleftharpoons$ Lanczos Method

- $\mathbf{b}_{PLS}^{(m)} = \mathbf{R}^{(m)}[(\mathbf{R}^{(m)})^T\mathbf{X}^T\mathbf{X}\mathbf{R}^{(m)}]^{-1}(\mathbf{R}^{(m)})^T\mathbf{X}^T\mathbf{y}$
- $\mathbf{R}^{(m)}$ - a matrix with orthonormal columns spanning Krylov space $\mathcal{K}^{(m)} = span\{\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}, \ldots, (\mathbf{X}^T\mathbf{X})^{m-1}\mathbf{X}^T\mathbf{y}\}$
  $\mathbf{W}^{(m)} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m]$ is such a candidate
- $\mathbf{Z}^{(m)} = (\mathbf{R}^{(m)})^T\mathbf{X}^T\mathbf{X}\mathbf{R}^{(m)}$ is a tridiagonal matrix
- Lanczos method approximate extremal eigenvalues of $\mathbf{X}^T\mathbf{X}$ by constructing a sequence of $\mathbf{Z}^{(m)}$; columns of $\mathbf{R}^{(m)}$ are given by a Gram-Schimdt orthonormalization of the first m columns of $\mathcal{K}^{(m)}$

PLS1 $\rightleftharpoons$ Conjugate Gradients (CG)

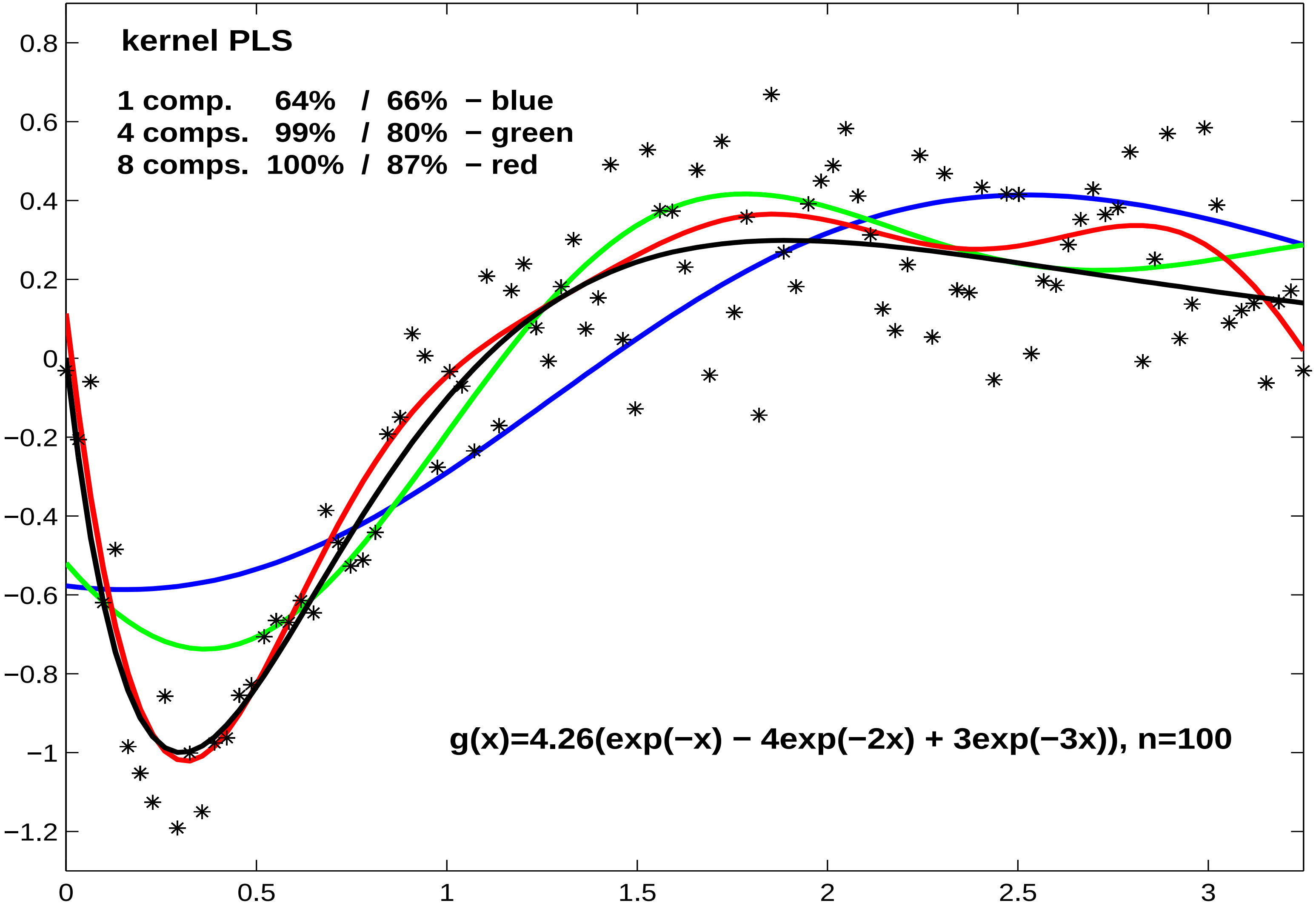
- CG - solves a system of linear equations $\mathbf{Af} = \mathbf{g}$ by minimization of the quadratic form $\frac{1}{2}\mathbf{f}^T\mathbf{Af} - \mathbf{g}^T\mathbf{f}$ ($\mathbf{A}$ positive semidefinite)
- for any $\mathbf{f}_0$, the sequence $\mathbf{f}_j$, iterates to the solution $\mathbf{f} = \mathbf{A}^{-}\mathbf{g}$ in $p = rank(\mathbf{A})$ steps
- the connection between CG and Lanczos method known (Hestens & Stiefel'52; Lanczos'50)
- if $\mathbf{A} = \mathbf{X}^T\mathbf{X}$; $\mathbf{g} = \mathbf{X}^T\mathbf{y}$ & $\mathbf{f}_0 = \mathbf{0}$ then $\mathbf{b}_{PLS}^{(m)} \rightleftharpoons \mathbf{f}_m$

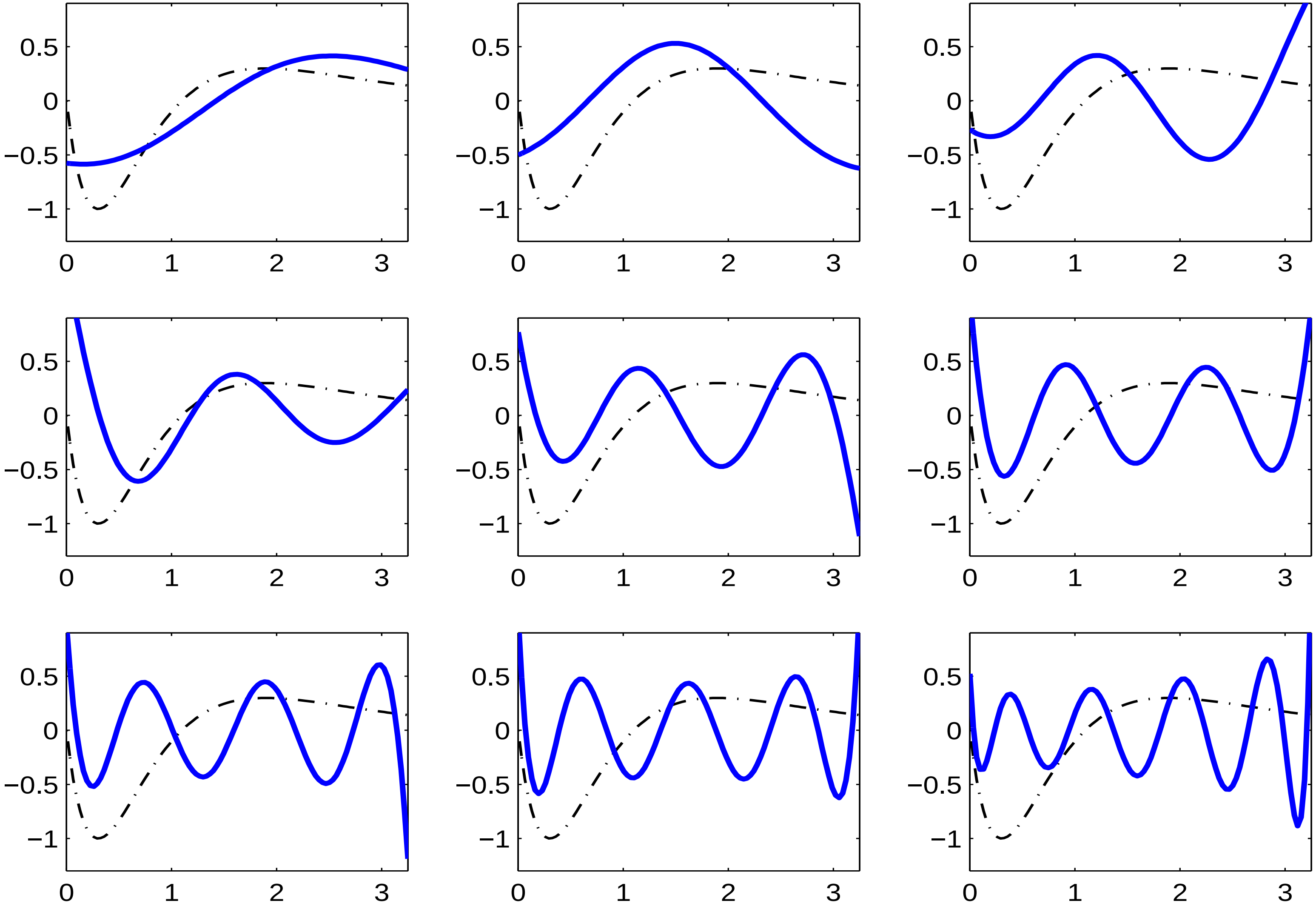
Kernel PLS Regression

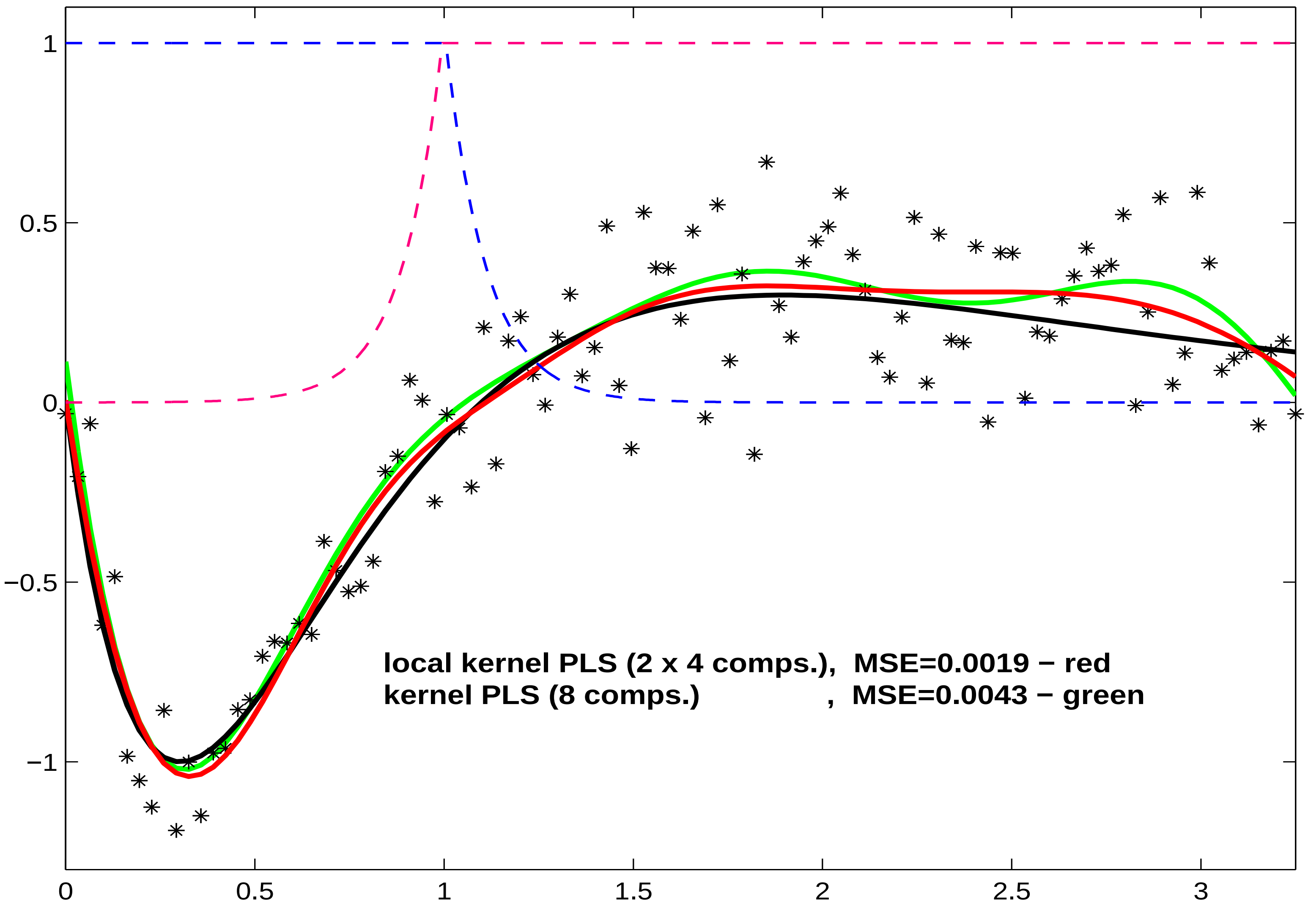
- linear PLS regression in a feature space $\mathcal{F}$
- kernel trick: $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$
  where $\mathbf{\Phi}$ is the $(n \times L)$ matrix of the mapped input data:
  $\Phi : \mathbf{x} \rightarrow \mathbf{\Phi}(\mathbf{x}) \in \mathcal{F}$
- nonlinear kernel-based PLS:
  $\mathbf{X}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t} \Rightarrow \quad \mathbf{K}\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t}$
  $\mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t}$

  or

  iterative kernel-based NIPALS algorithm


kernel PLS
1 comp. 64% / 66% – blue
4 comps. 99% / 80% – green
8 comps. 100% / 87% – red
g(x)=4.26(exp(−x) − 4exp(−2x) + 3exp(−3x)), n=100
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
0
0.5
1
1.5
2
2.5
3


0.5
0
−0.5
−1
0
1
2
3


local kernel PLS (2 x 4 comps.), MSE=0.0019 − red
kernel PLS (8 comps.) , MSE=0.0043 − green
1
0.5
0
−0.5
−1
0
0.5
1
1.5
2
2.5
3

"The Peculiar Shrinkage Properties" of PLS1

(Frank & Friedman'93, Butler & Denham'00, Lingjaerde & Christophersen'00, Krämer'04)

- assume: $\mathbf{y} = \mathbf{X}\mathbf{b} + \epsilon$

  $\mathbf{y}$ an $(n \times 1)$ response vector

  $\mathbf{X}$ an $(n \times N)$ design matrix

  $\mathbf{b}$ an unknown $(N \times 1)$ parameter vector

  ϵ an $(n \times 1)$ vector of noise, iid elements $\sim \mathcal{N}(0, \sigma^2)$

  $\mathbf{y}, \mathbf{X}$ centered, i.e. $\mathbf{1}_n^T \mathbf{Y} = 0$ and $\mathbf{1}_n^T \mathbf{X} = \mathbf{0}_N$,

  $rank(\mathbf{X}) = p \leq \min(n-1, N)$

  svd$(\mathbf{X}) = \mathbf{U}\mathbf{D}\mathbf{V}^T$; δ_i - singular values

  $\mathbf{X}^T\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$, $\lambda_i = \delta_i^2$

Ordinary Least Squares (OLS)

- $min_{\mathbf{b}}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|_2 \Longrightarrow \hat{\mathbf{b}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y} = \mathbf{V}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{Y}$
  $\hat{\mathbf{b}}_{OLS} = \sum_{i=1}^{p}\lambda_i^{-1/2}(\mathbf{u}_i^T\mathbf{y})\mathbf{v}_i = \sum_{i=1}^{p}\hat{\mathbf{b}}_i$
- $\hat{\mathbf{b}}_{OLS}$ belongs to the class of linear estimators $\hat{\mathbf{z}} = \mathbf{L}\mathbf{y}$
  $E(\hat{\mathbf{z}}) = \mathbf{L}\mathbf{X}\mathbf{z}$
  $var(\hat{\mathbf{z}}) = \sigma^2 trace(\mathbf{L}\mathbf{L}^T)$
- $E(\hat{\mathbf{b}}_{OLS}) = \mathbf{b}$
  $var(\hat{\mathbf{b}}_{OLS}) = E[(\hat{\mathbf{b}}_{OLS}-\mathbf{b})^T(\hat{\mathbf{b}}_{OLS}-\mathbf{b})] = \sigma^2 trace(\mathbf{X}^T\mathbf{X})^{-} =$
  $= \sigma^2\sum_{i=1}^{m}\frac{1}{\lambda_i}$
- $MSE(\hat{\mathbf{z}}) = (E(\hat{\mathbf{z}})-\mathbf{z})^T(E(\hat{\mathbf{z}})-\mathbf{z}) + E[(\hat{\mathbf{z}}-E(\hat{\mathbf{z}}))^T(\hat{\mathbf{z}}-E(\hat{\mathbf{z}}))]$
  $\equiv bias^2(\hat{\mathbf{z}}) + var(\hat{\mathbf{z}})$
- if $\|\hat{\mathbf{z}}_1\|_2 \leq \|\hat{\mathbf{z}}_2\|_2 \Rightarrow var(\hat{\mathbf{z}}_1) \leq var(\hat{\mathbf{z}}_2)$

Shrinkage Estimators

- $\hat{\mathbf{b}}_{shr} = \sum_{i=1}^{p} f(\lambda_i)\lambda_i^{-1/2}(\mathbf{u}_i^T\mathbf{y})\mathbf{v}_i = \sum_{i=1}^{p} f(\lambda_i)\hat{\mathbf{b}}_i$

  $\hat{\mathbf{b}}_i$ – the component of $\hat{\mathbf{b}}_{OLS}$ along $\mathbf{v}_i$

- linear shrinkage estimators

$$MSE(\hat{\mathbf{b}}_{shr}) = \sum_{i=1}^{p}(f(\lambda_i)-1)^2(\mathbf{v}_i^T\mathbf{b})^2 + \sigma^2\sum_{i=1}^{p} f(\lambda_i)^2/\lambda_i$$

(Generalized) Ridge Regression

$$f(\lambda_i) = \frac{\lambda_i}{\lambda_i + \gamma_i}\ , \quad \gamma_i \; - \; \text{regularization term along } \mathbf{v}_i$$

Principal Components Regression (PCR)

$$f(\lambda_i) = \begin{cases} 1 & : \quad \text{principal component along } \mathbf{v}_i \text{ included} \\ 0 & : \quad \text{otherwise} \end{cases}$$

PLS Regression (PLS1)

- $\hat{\mathbf{b}}_{PLS}^{(m)} = \sum_{i=1}^{p} f^{(m)}(\lambda_i)\hat{\mathbf{b}}_i$
- $\hat{\mathbf{b}}_{PLS}^{(m)}$ is not a linear estimator
- PLS shrinks:
  $\|\hat{\mathbf{b}}_{PLS}^{(1)}\|_2 \leq \|\hat{\mathbf{b}}_{PLS}^{(2)}\|_2 \leq \ldots \leq \|\hat{\mathbf{b}}_{PLS}^{(p)}\|_2 = \|\hat{\mathbf{b}}_{OLS}\|_2$
- PLS fits closer to OLS then PCR:
  $\mathcal{R}^2(\hat{\mathbf{y}}_{OLS}, \hat{\mathbf{y}}_{PLS}^{(m)}) \geq \mathcal{R}^2(\hat{\mathbf{y}}_{OLS}, \hat{\mathbf{y}}_{PCR}^{(m)})$
  ($\mathcal{R}^2(.,.)$ - squared correlation)

PLS Shrinkage Factors $f^{(m)}(\lambda_i)$

- $$f^{(m)}(\lambda_i) = 1 - \prod_{j=1}^{m}(1 - \frac{\lambda_i}{\mu_j^{(m)}}) \ , \quad i = 1, \ldots, p$$

  $\mu_1^{(m)} \geq \ldots \geq \mu_m^{(m)}$ the eigenvalues (Ritz values) of $(\mathbf{R}^{(m)})^T\mathbf{X}^T\mathbf{X}\mathbf{R}^{(m)}$

- $\mathbf{R}^{(m)}$ - a matrix with orthonormal columns spanning Krylov space $\mathcal{K}^{(m)} = span\{\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}, \ldots, (\mathbf{X}^T\mathbf{X})^{m-1}\mathbf{X}^T\mathbf{y}\}$
  $\mathbf{W}^{(m)} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m]$ is such a candidate

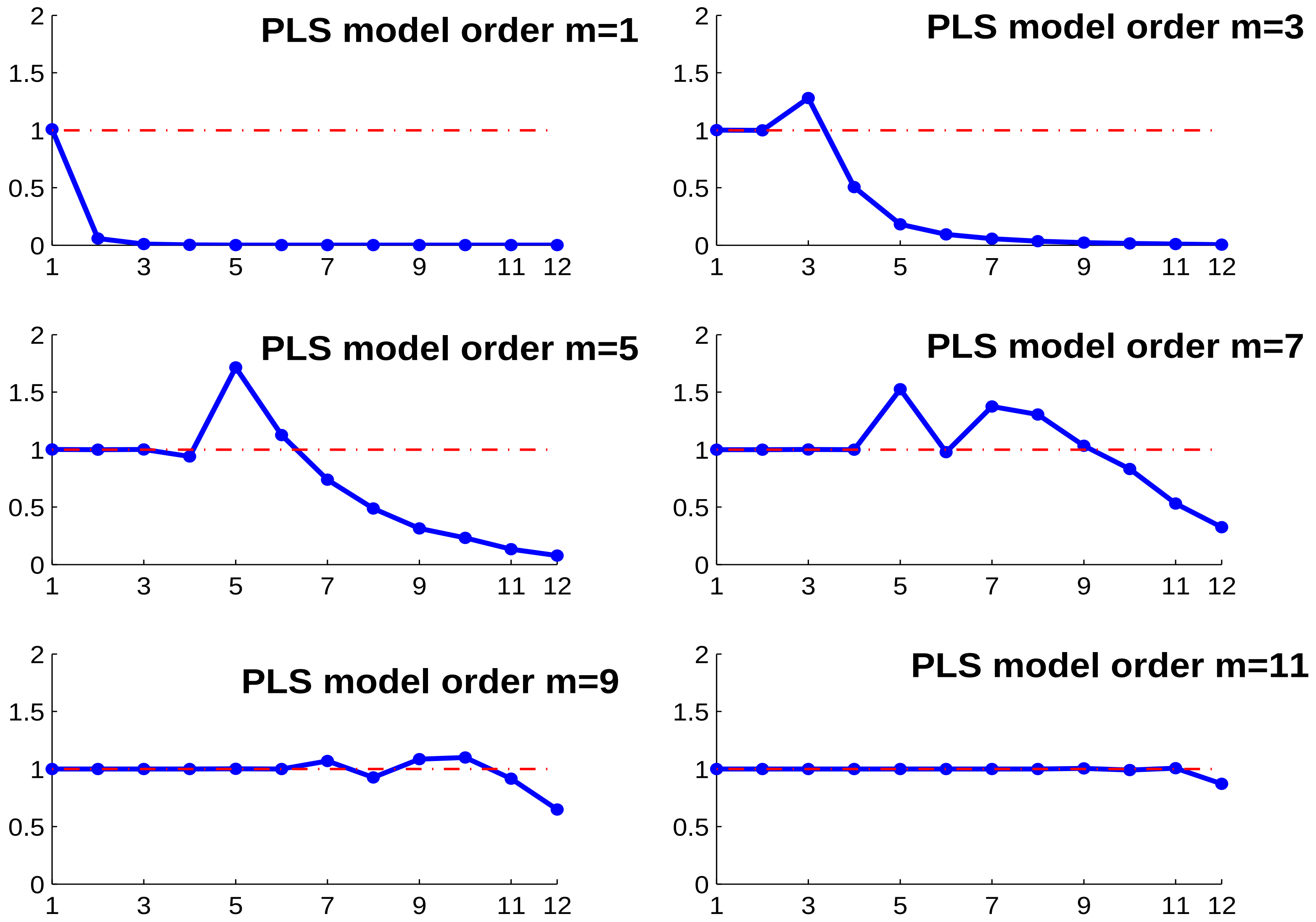
Fundamental Properties of $f^{(m)}(\lambda_i)$

- $f^{(m)}(\lambda_i)$ depends non-linearly on $\mathbf{y}$
- $f^{(m)}(\lambda_i) > 1$ may occur
- $f^{(m)}(\lambda_p) \leq 1$ for all m
- $f^{(m)}(\lambda_1) \geq 1$ for all $m = 1, 3, 5, \ldots$
- $f^{(m)}(\lambda_1) \leq 1$ for all $m = 2, 4, 6, \ldots$
- for $m < M$ (M - number of distinct eigenvalues of $\mathbf{X}^T\mathbf{X}$)

  (i) at least $(m+1)/2$ shrink. factors satisfy $f^{(m)}(\lambda_i) \geq 1$

  (ii) at least $(m/2)+1$ shrink. factors satisfy $f^{(m)}(\lambda_i) \leq 1$

  (iii) there exist an $i \geq m$ such that $f^{(m)}(\lambda_i) \geq 1$


PLS model order m=1
PLS model order m=3
PLS model order m=5
PLS model order m=7
PLS model order m=9
PLS model order m=11
0
0.5
1
1.5
2
1
3
5
7
9
11
12

Multiple Multivariate PLS Regression

- prediction when a high degree of correlation among the variables in both the predictor and response spaces exist
- PLS2 is inherently designed to deal with several response variables, however, almost none theoretical understanding of the properties of such model exist
- the curds & whey procedure (C&W) (Breiman & Friedman'97): the use of CCA between predictors and responses to decorrelate response variables $\Rightarrow$ univariate (shrinkage) regression on decorrelated responses
- experimental evidence exists that C&W in the PLS2 framework may improve prediction accuracies (Xu & Massart'03)

Selection of Variables (PLS1) - CovProc

- $\mathbf{t} = \mathbf{X}\mathbf{w}$; explained variance (fit) associated with $\mathbf{t}$ is $r^2 = (\mathbf{y}^T\mathbf{t})^2/(\mathbf{t}^T\mathbf{t})$
- let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ and weight vector $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2]$:
  $(\mathbf{y}^T\mathbf{t})^2 = ((\mathbf{y}^T\mathbf{X}_1\mathbf{w}_1) + (\mathbf{y}^T\mathbf{X}_2\mathbf{w}_2))$
  $\mathbf{t}^T\mathbf{t} = \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{w}_1 + 2\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_1\mathbf{w}_1 + \mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{w}_2$
- problem: large $(\mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_2\mathbf{w}_2)$ can spoil good fit given by large $(\mathbf{y}^T\mathbf{X}_1\mathbf{w}_1)$; e.g large amount of small components in $\mathbf{w}$
- (i) compute $\mathbf{w}$ using $\mathbf{X}$
  (ii) sort $\mathbf{x}_i$ using $abs(\mathbf{w})$
  (iii) compute r^2 and/or cross-validate sub-models
  (iv) compute new PLS model $(\mathbf{w}, \mathbf{t}, \ldots)$ using selected $\mathbf{x}_i$

PLS Discrimination/Classification

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

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

- orthonormalized PLS

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

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

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

  $$\mathbf{Hw} = \lambda\mathbf{w}$$

- CCA, Fisher's LDA

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

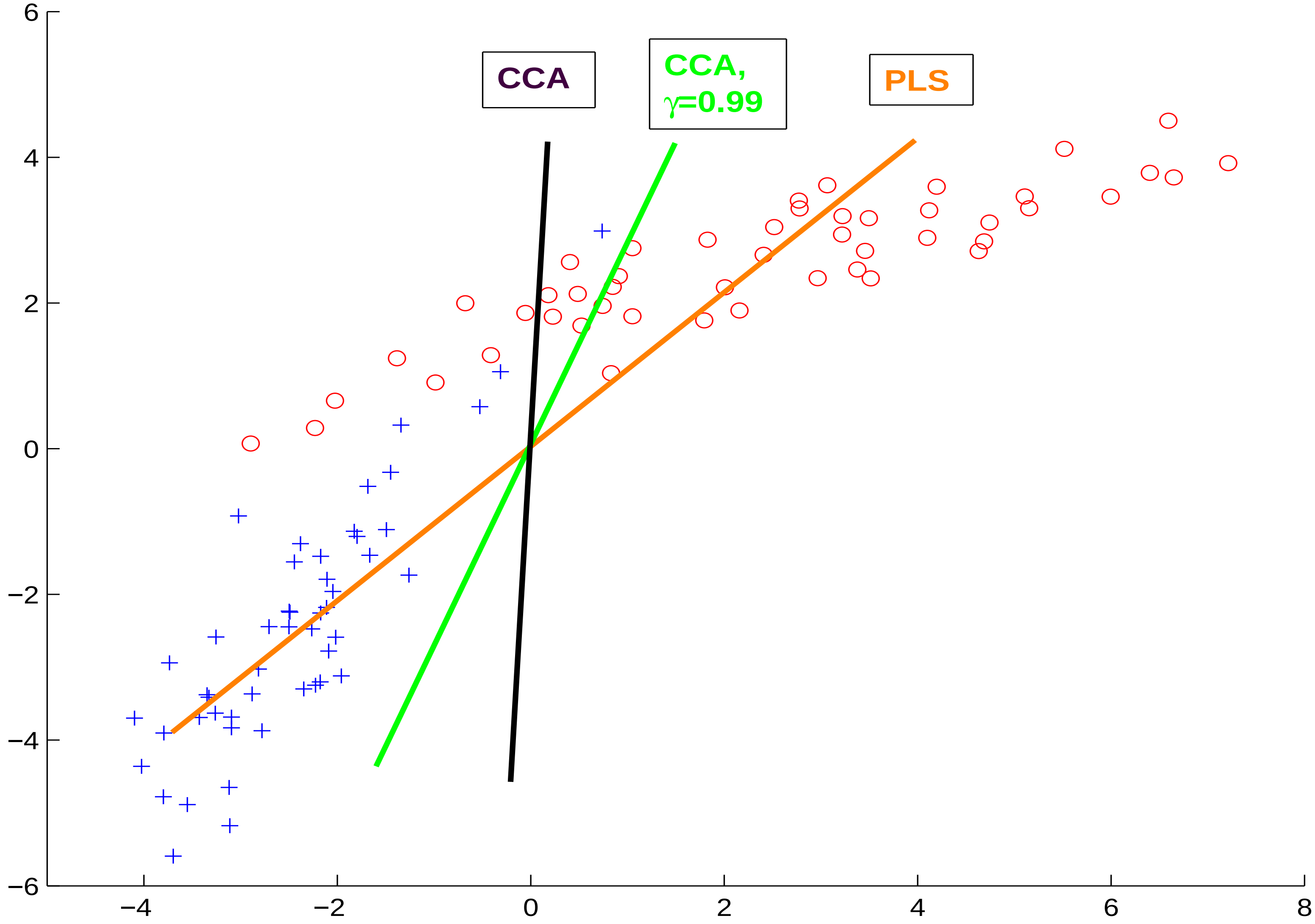
  $$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{Xa} = \lambda\mathbf{a}$$

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

Canonical Ridge Analysis - CCA $\rightleftharpoons$ PLS

$$([1-\gamma_X]\mathbf{X}^T\mathbf{X}+\gamma_X\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}([1-\gamma_Y]\mathbf{Y}^T\mathbf{Y}+\gamma_X\mathbf{I})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{w}=\lambda\mathbf{w}$$

- CCA: $\gamma_X = 0, \; \gamma_Y = 0$
- PLS: $\gamma_X = 1, \; \gamma_Y = 1$
- Orthonormalized PLS: $\gamma_X = 1, \; \gamma_Y = 0$ or $\gamma_X = 0, \; \gamma_Y = 1$
- Ridge Regression, Regularized FDA or CCA:
  $\gamma_X \in (0,1), \; \mathbf{Y} \in \mathcal{R}$


CCA
CCA,
γ=0.99
PLS
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8

Kernel PLS Discrimination

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

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

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

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)

Kernel PLS Pseudocode ($\mathbf{Y} \subseteq \mathcal{R}$)

1. kernel PLS score vectors extraction

   compute $\mathbf{K}$ - centered Gram matrix

   set $\mathbf{K}_{res} = \mathbf{K}$, m - the number of score vectors

   for $i = 1$ to m

   $\mathbf{t}_i = \mathbf{K}_{res}\mathbf{Y}$

   $\|\mathbf{t}_i\| \rightarrow 1$

   $\mathbf{u}_i = \mathbf{Y}(\mathbf{Y}^T\mathbf{t}_i)$

   $\mathbf{K}_{res} \leftarrow \mathbf{K}_{res} - \mathbf{t}_i(\mathbf{t}_i^T\mathbf{K}_{res})$

   $\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t}_i(\mathbf{t}_i^T\mathbf{Y})$

   end

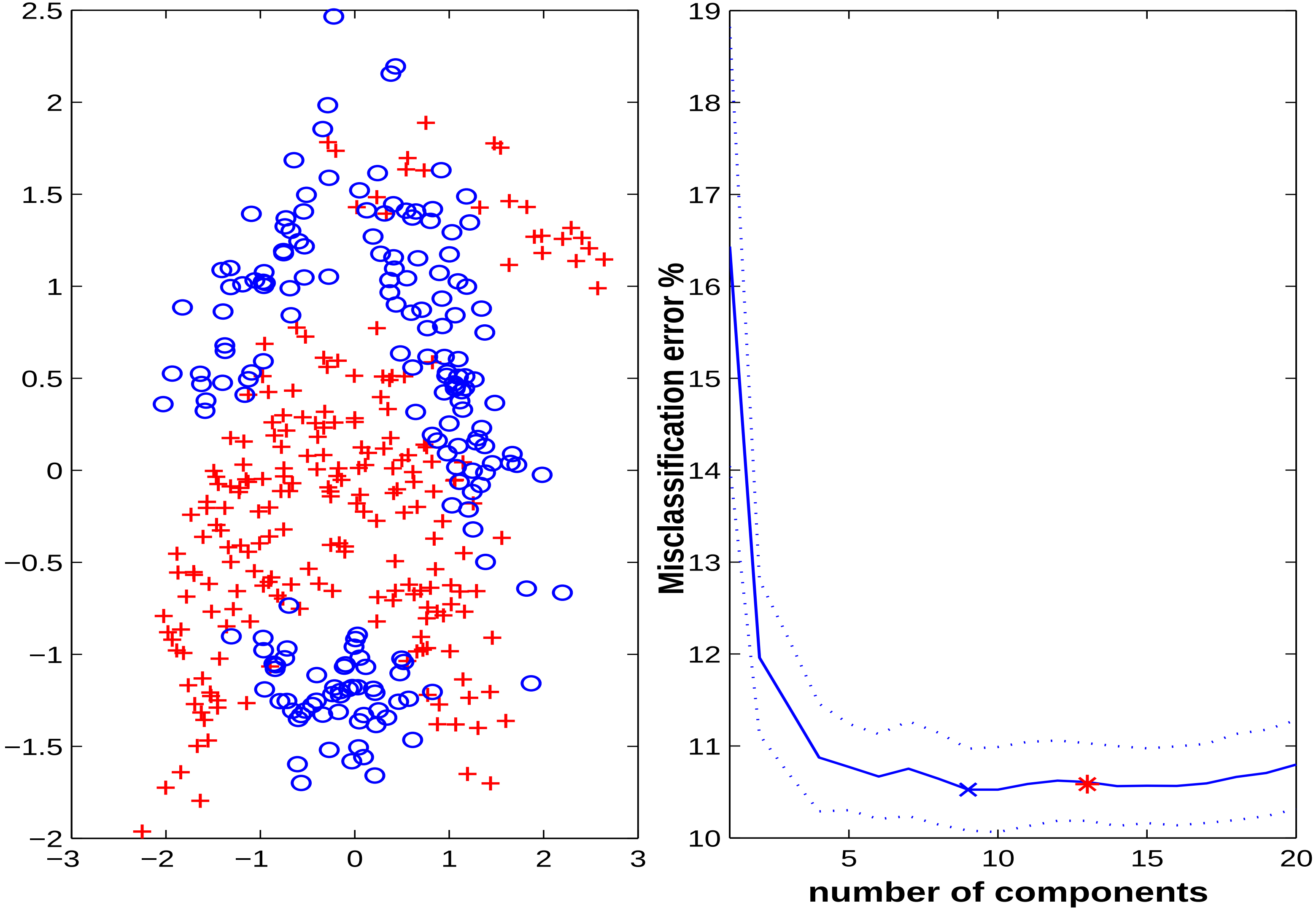
2. projection of test samples

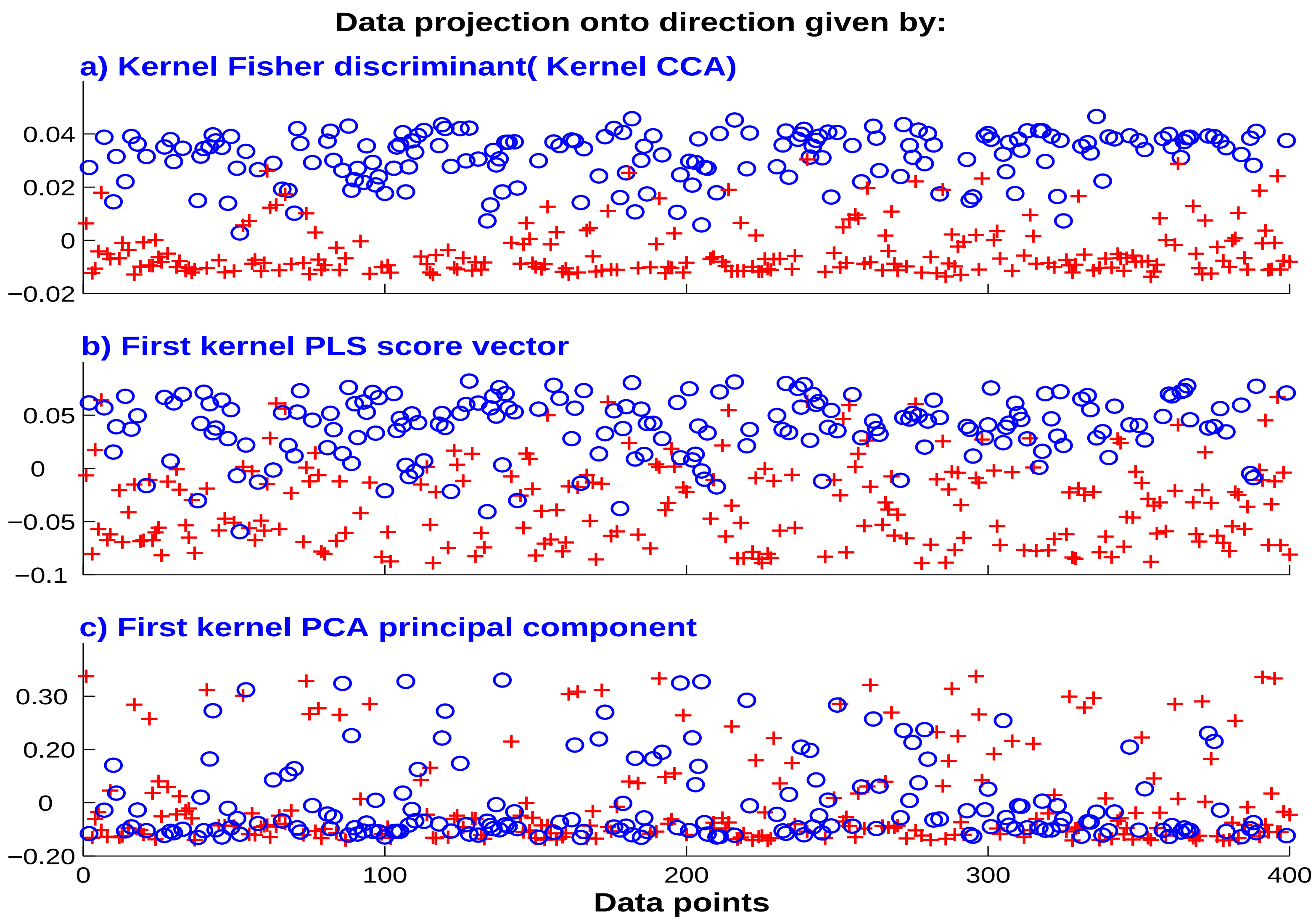
   $\mathbf{T}_t = \mathbf{K}_t\mathbf{U}(\mathbf{T}^T\mathbf{K}\mathbf{U})^{-1}$; ($\mathbf{K}_t$ - test set Gram matrix)

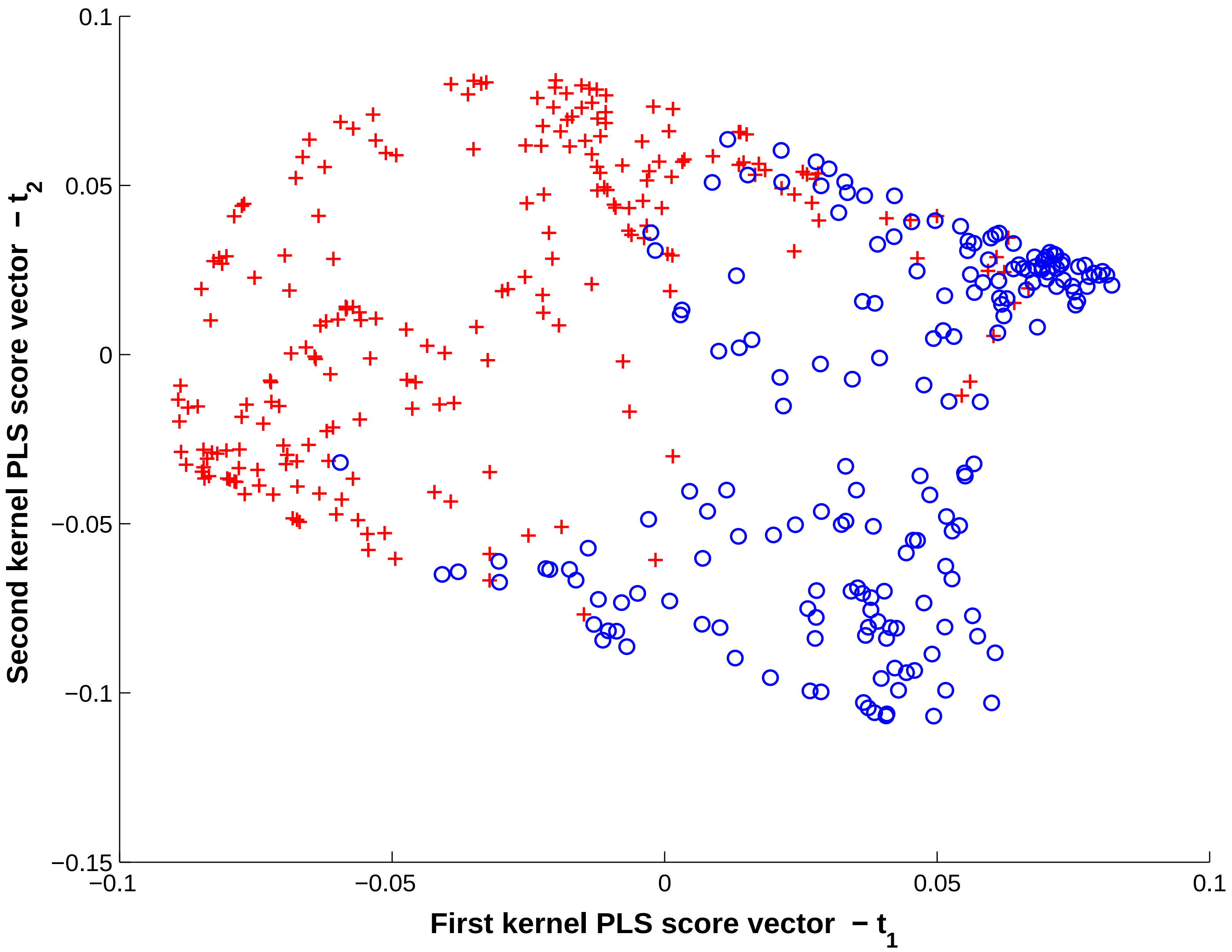
Experiments - Classification

- 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)
- cognitive fatigue estimation

Banana data set

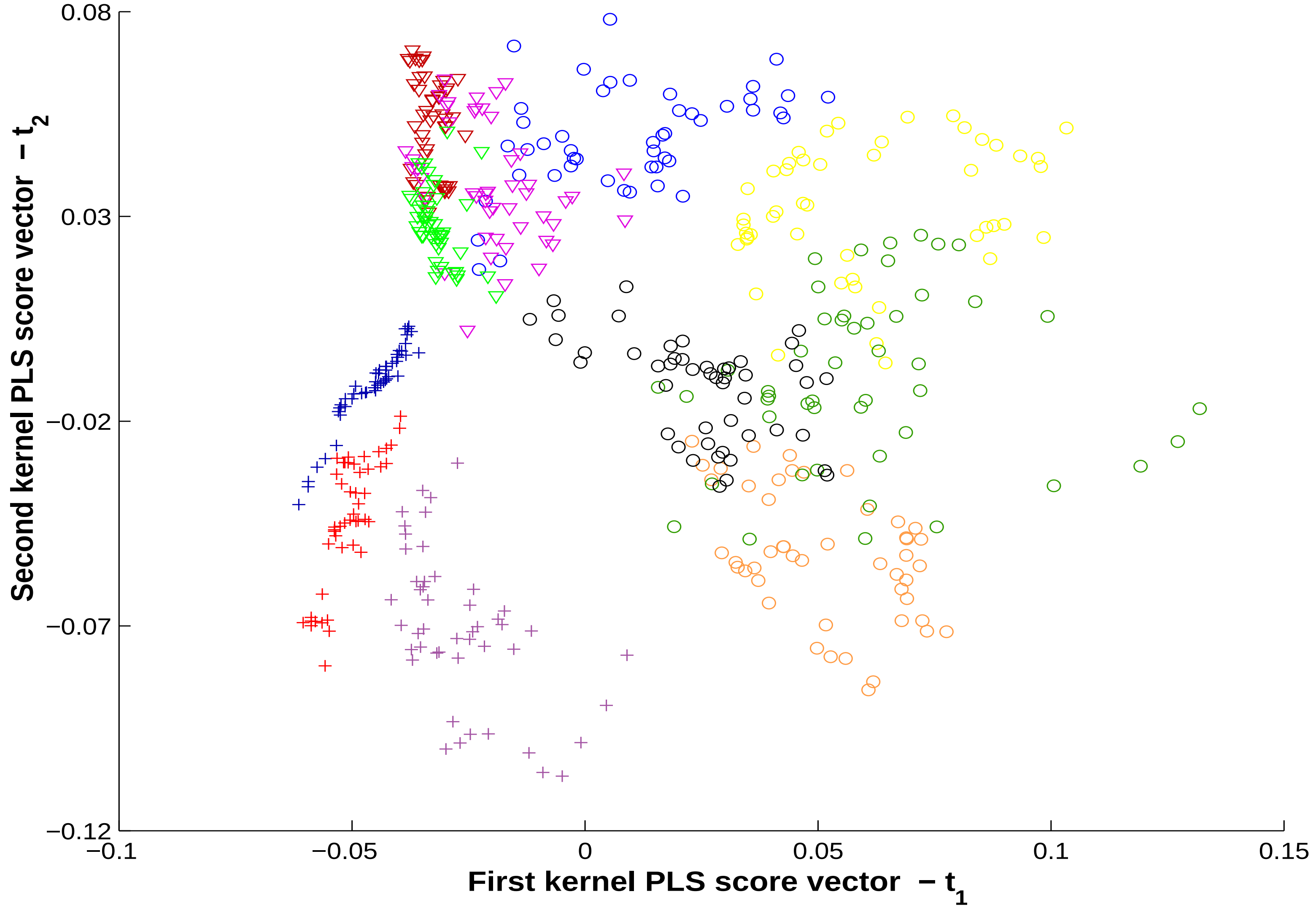



Data projection onto direction given by:
a) Kernel Fisher discriminant(Kernel CCA)
0.04
0.02
0
−0.02
b) First kernel PLS score vector
0.05
0
−0.05
−0.1
c) First kernel PCA principal component
0.30
0.20
0
−0.20
0
100
200
300
400
Data points

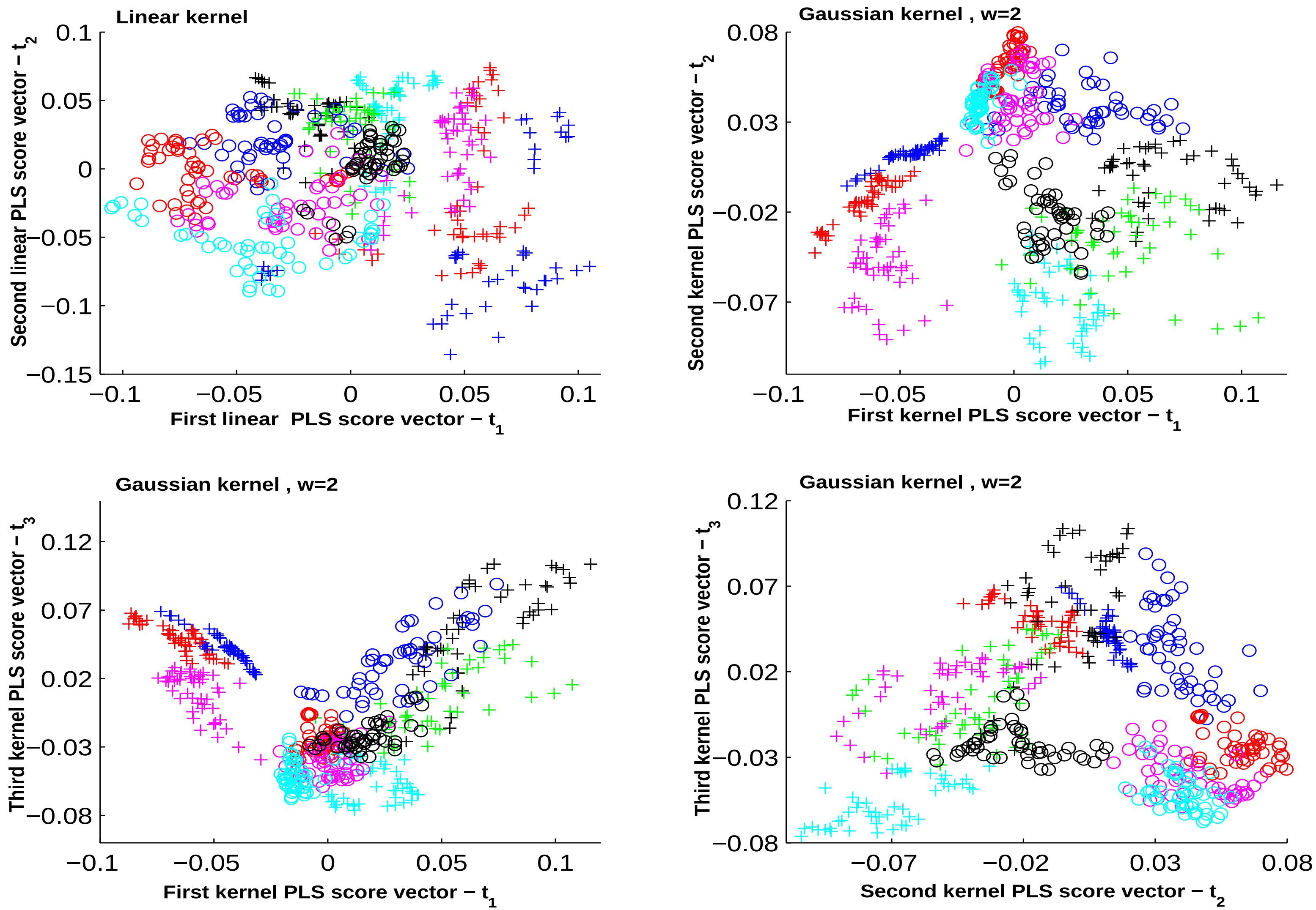

0.1
0.05
0
−0.05
−0.1
−0.15
−0.1
−0.05
0
0.05
0.1
Second kernel PLS score vector − t_2
First kernel PLS score vector − t_1

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



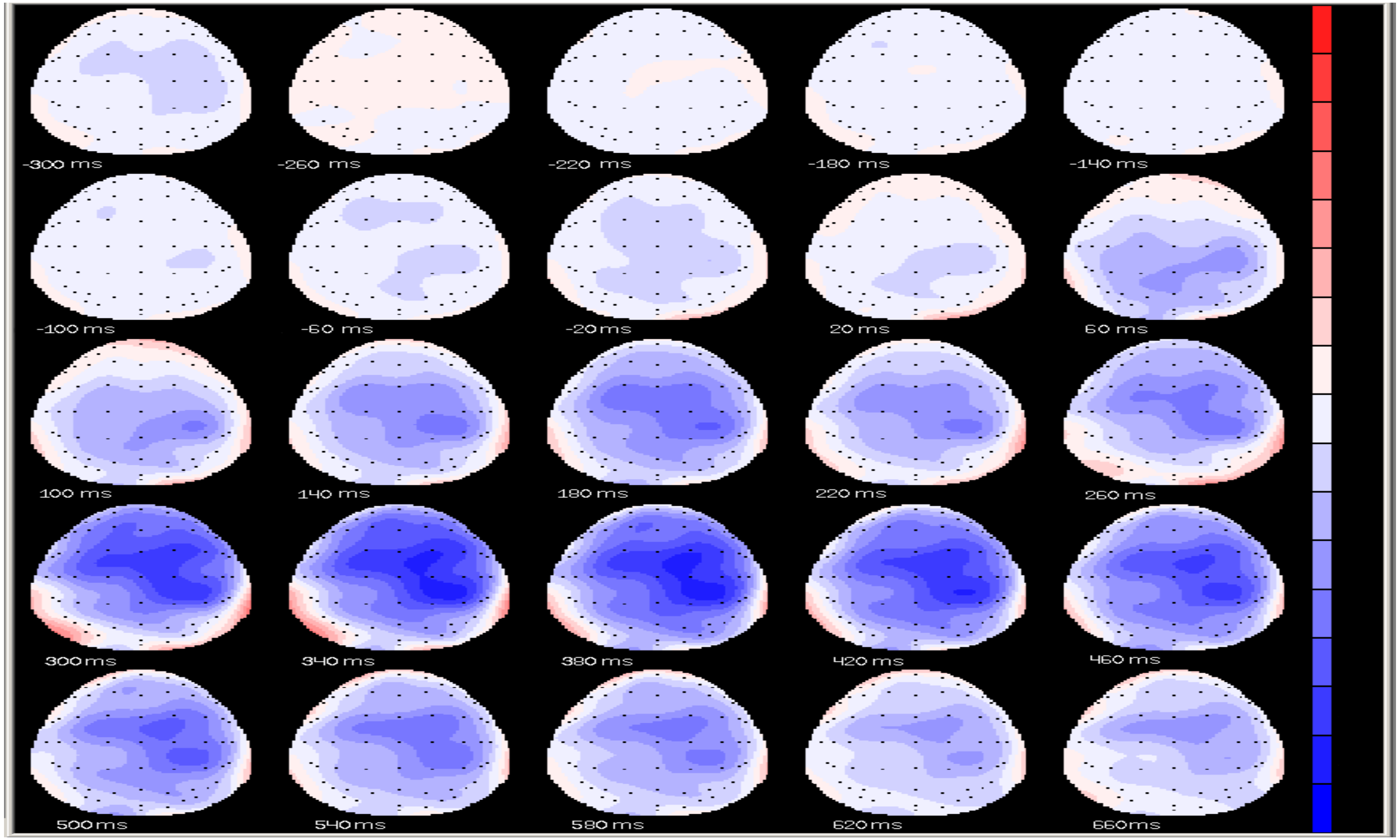
Vowel sounds data set: 11 classes, 10 predictors



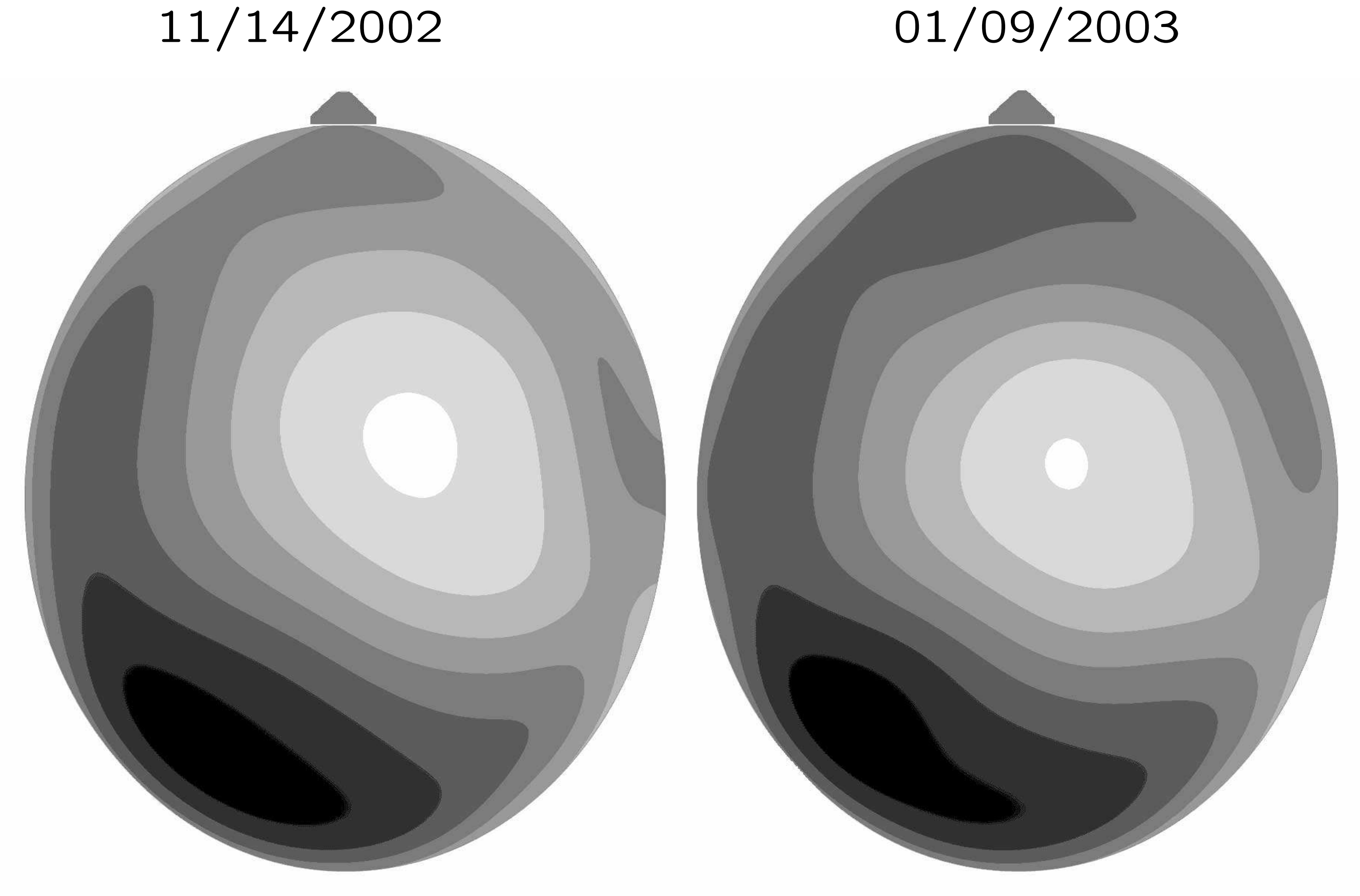
Method	Training Error	Testing Error
LDA	0.32	0.56
SVC (linear) - 1vs1	0.19	0.51
KPLS-SVC (linear) - 1vs1	0.16	0.47
FDA/MARS (df=2)	0.02	0.42
FDA/MARS (df=6,red. dim.)	0.13	0.39
SVC (gauss) - 1vs1	0.01	0.37
KPLS-SVC (gauss) - 1vs1	0.01	0.35
SVC (gauss, w $\leq$ 5) - 1vs1	0.002	0.29
KPLS-SVC (gauss, w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

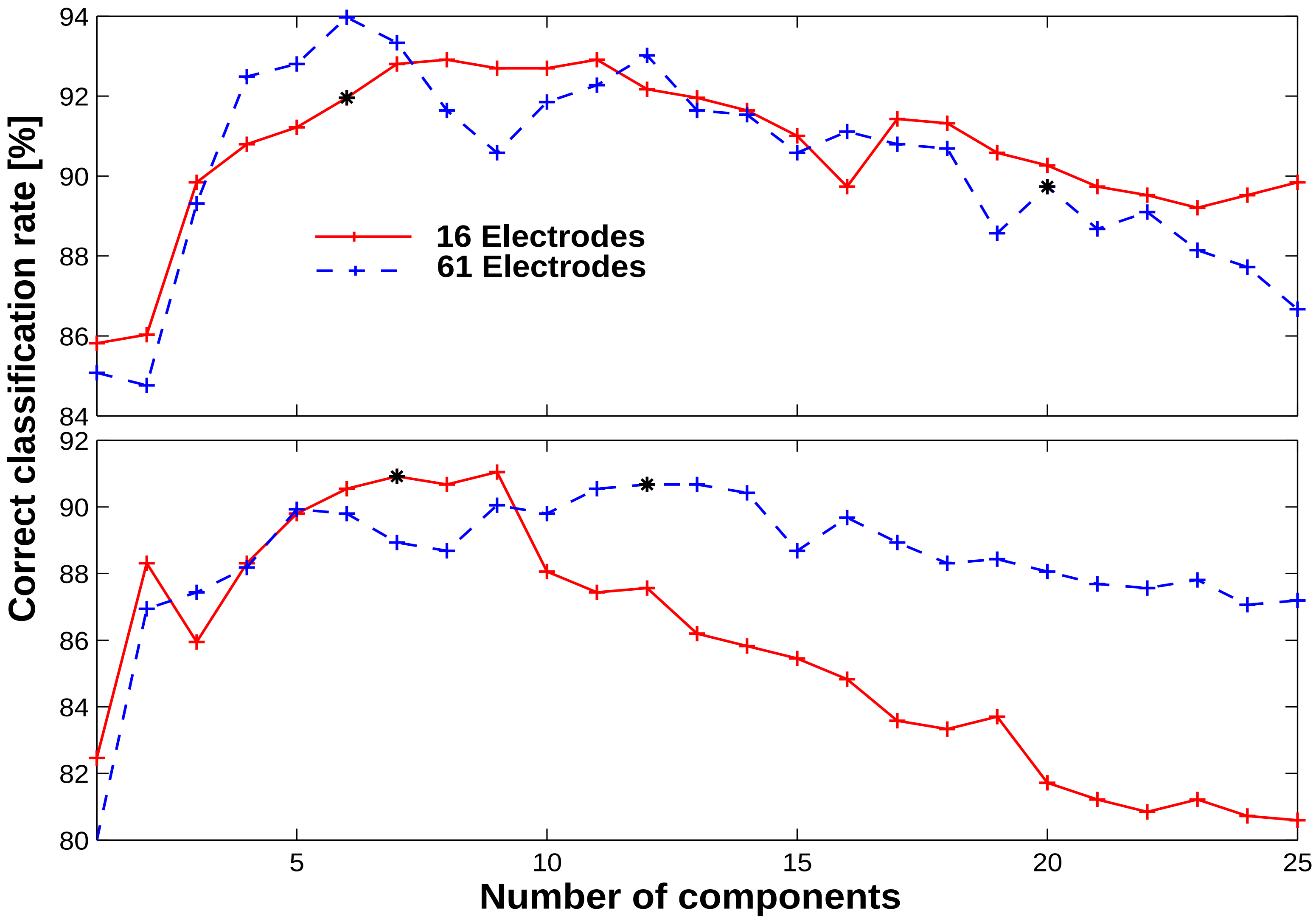
Finger movement periods vs. non-movement periods

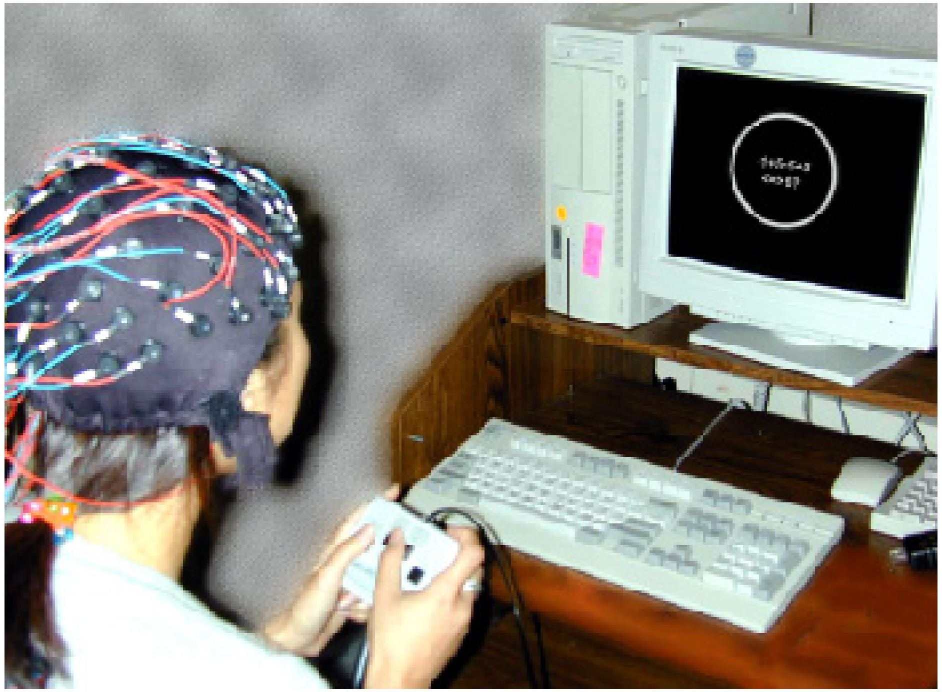
PLS-derived Spatio-temporal Filter - 01/09/2003

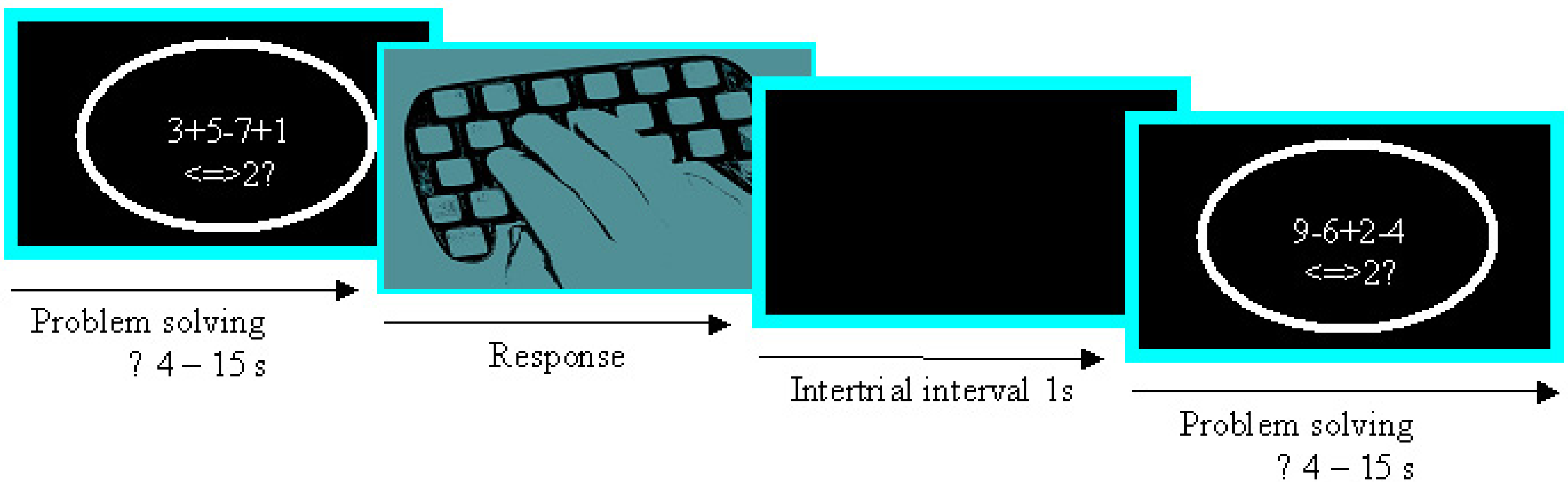


PLS-derived Spatio-temporal Filter
(370ms after button press)

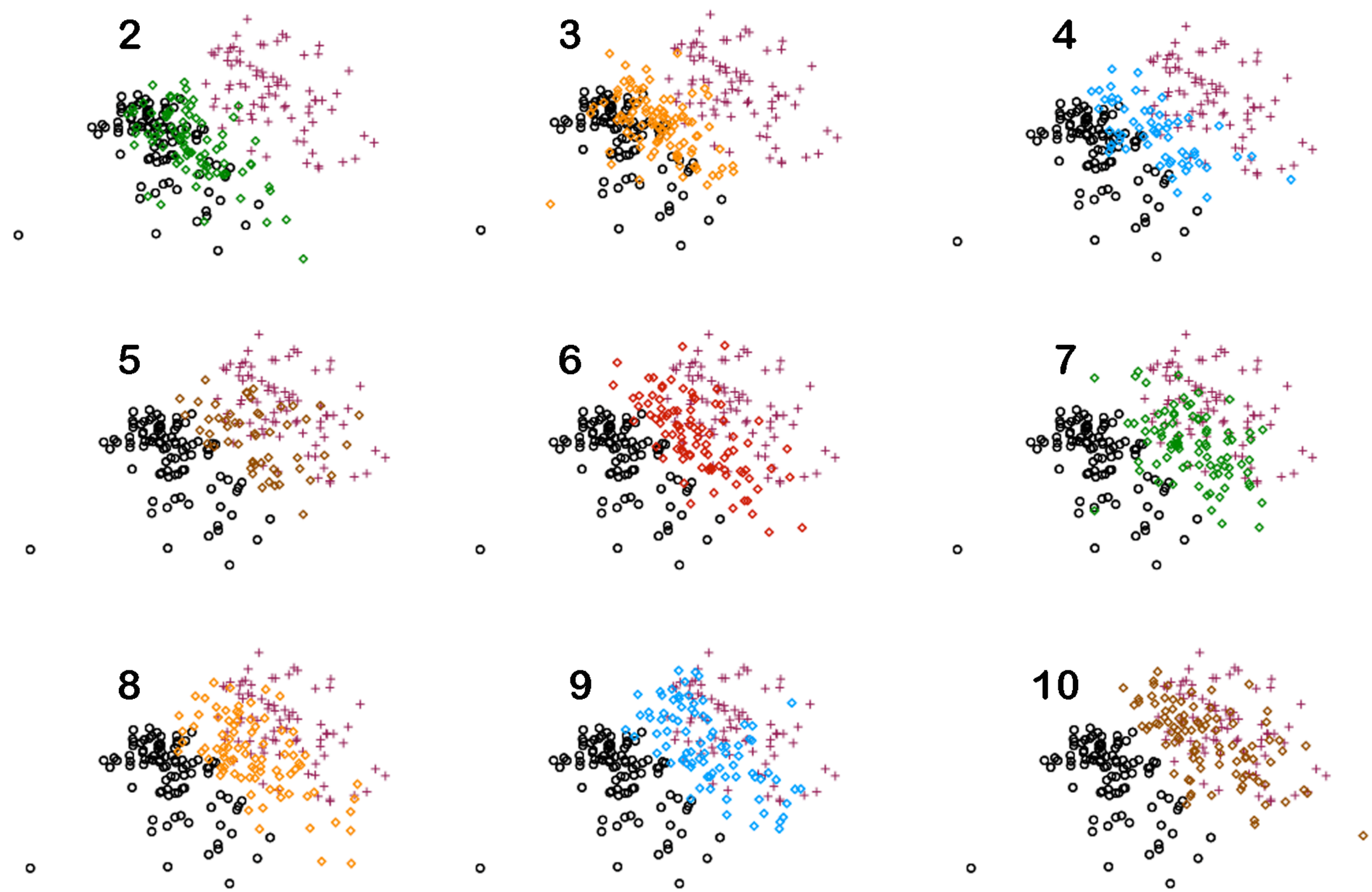



94
92
90
88
86
84
92
90
88
86
84
82
80
5
10
15
20
25
16 Electrodes
61 Electrodes
Correct classification rate [%]
Number of components




3+5-7+1
<=>2?
Problem solving
? 4 – 15 s
Response
Intertrial interval 1s
9-6+2-4
<=>2?
Problem solving
? 4 – 15 s

KPLS Scores (C1, C2) Predicted for EEG Epochs in the Intervening 15-minute Blocks



Kernel PLS Estimation of ERP - Regression

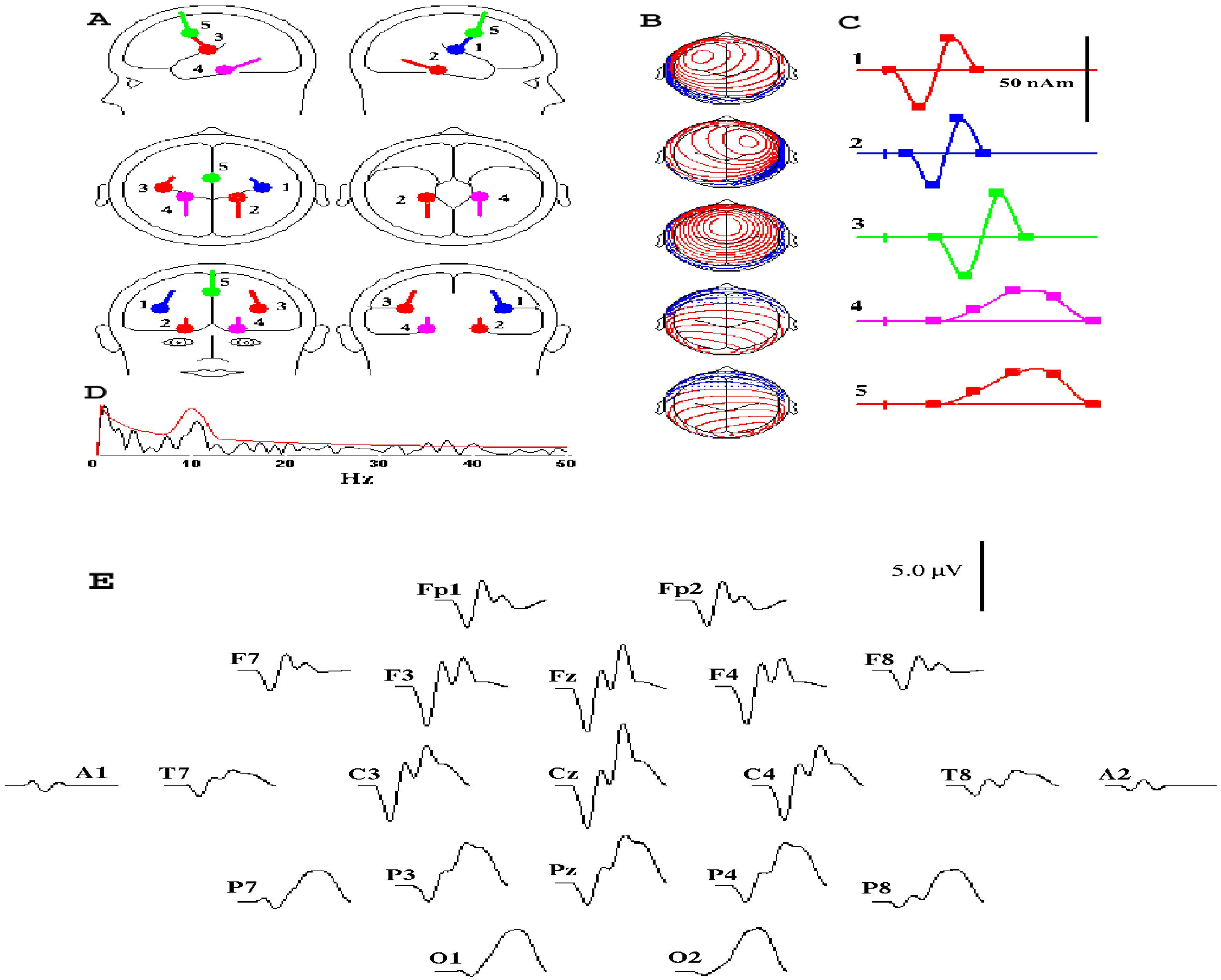
- Generated data:
  Event-Related Potentials (N1,P2,N2,P3)

  $$+$$

  relax state spatially distributed EEG signal + white Gaussian noise

- Real ERP data:
  ERPs recorded in an experiment of cognitive fatigue

Generation of ERPs using BESA software



Smoothing Splines

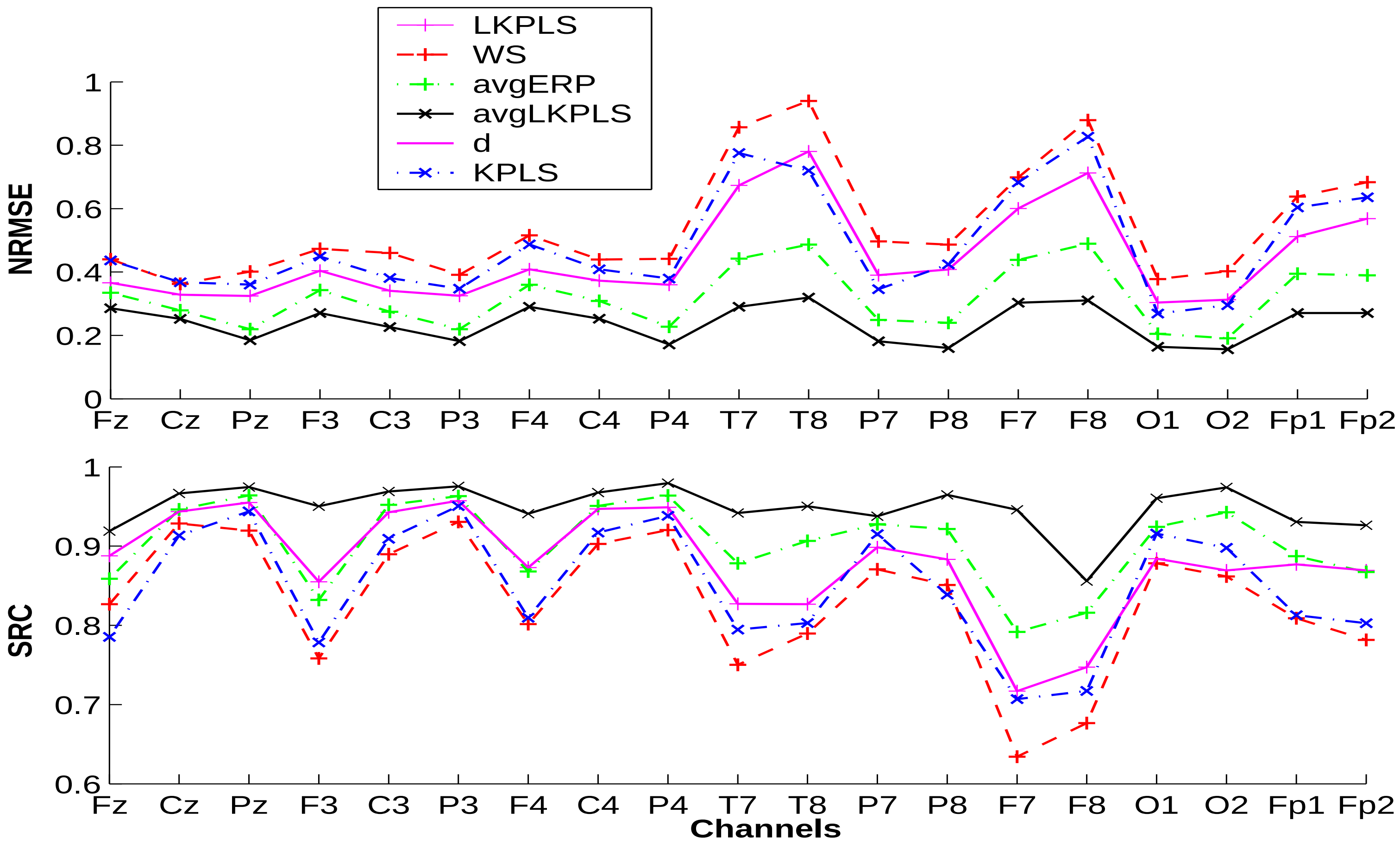
- $min_f(\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f^{(2)}(x))^2 dx \quad \lambda > 0 \Rightarrow$
  natural cubic splines with knots at $x_i\,; i = 1, \ldots, n$
- Complete basis $\rightarrow$ *shrink* the coefficients toward smoothing

Wavelet Smoothing

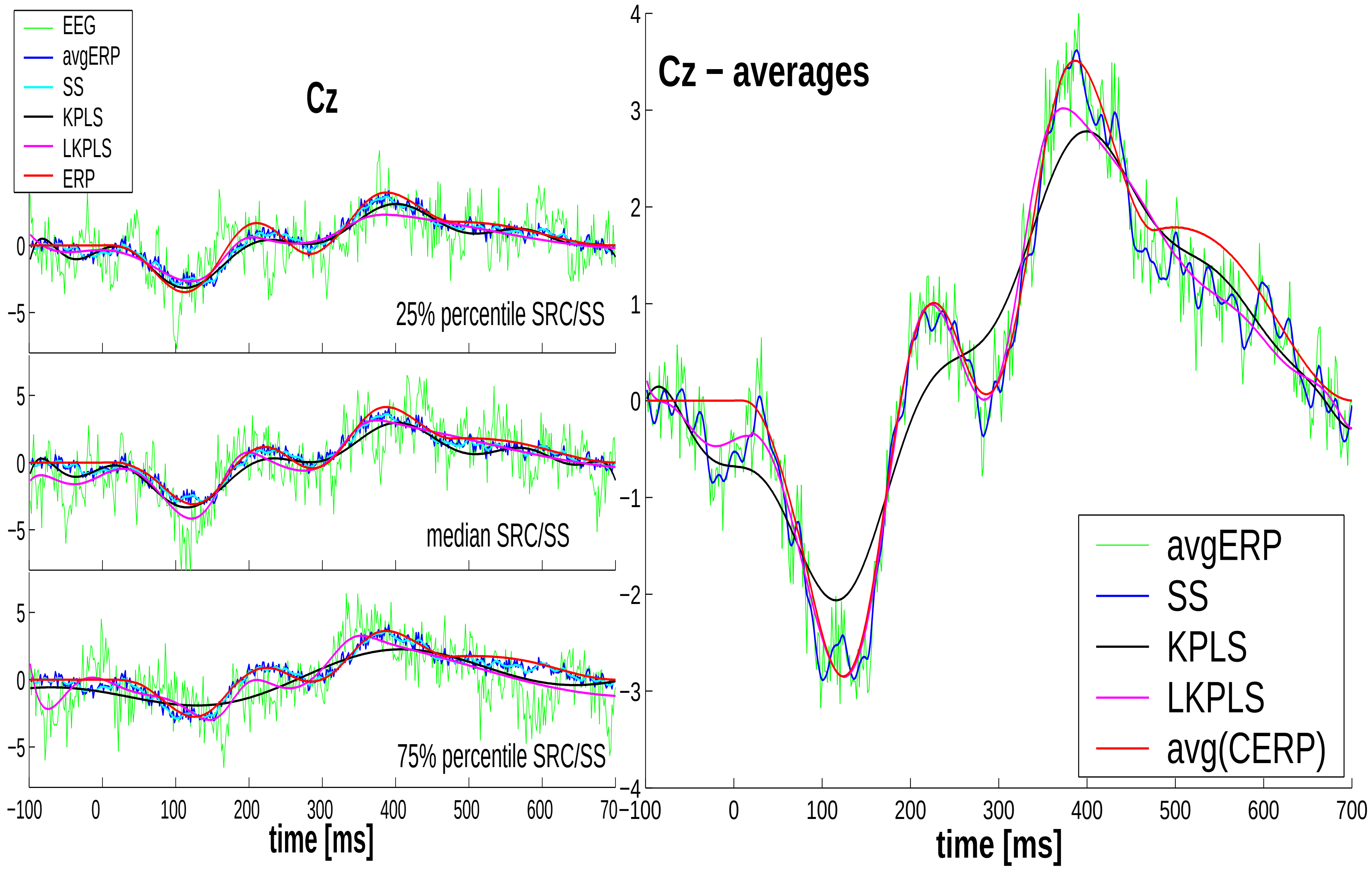
- Complete orthonormal basis $\rightarrow$ *shrink* and *select* the coefficients toward a **sparse** representation
- Wavelet basis is *localized in time and frequency*

Correlated Noise Estimate

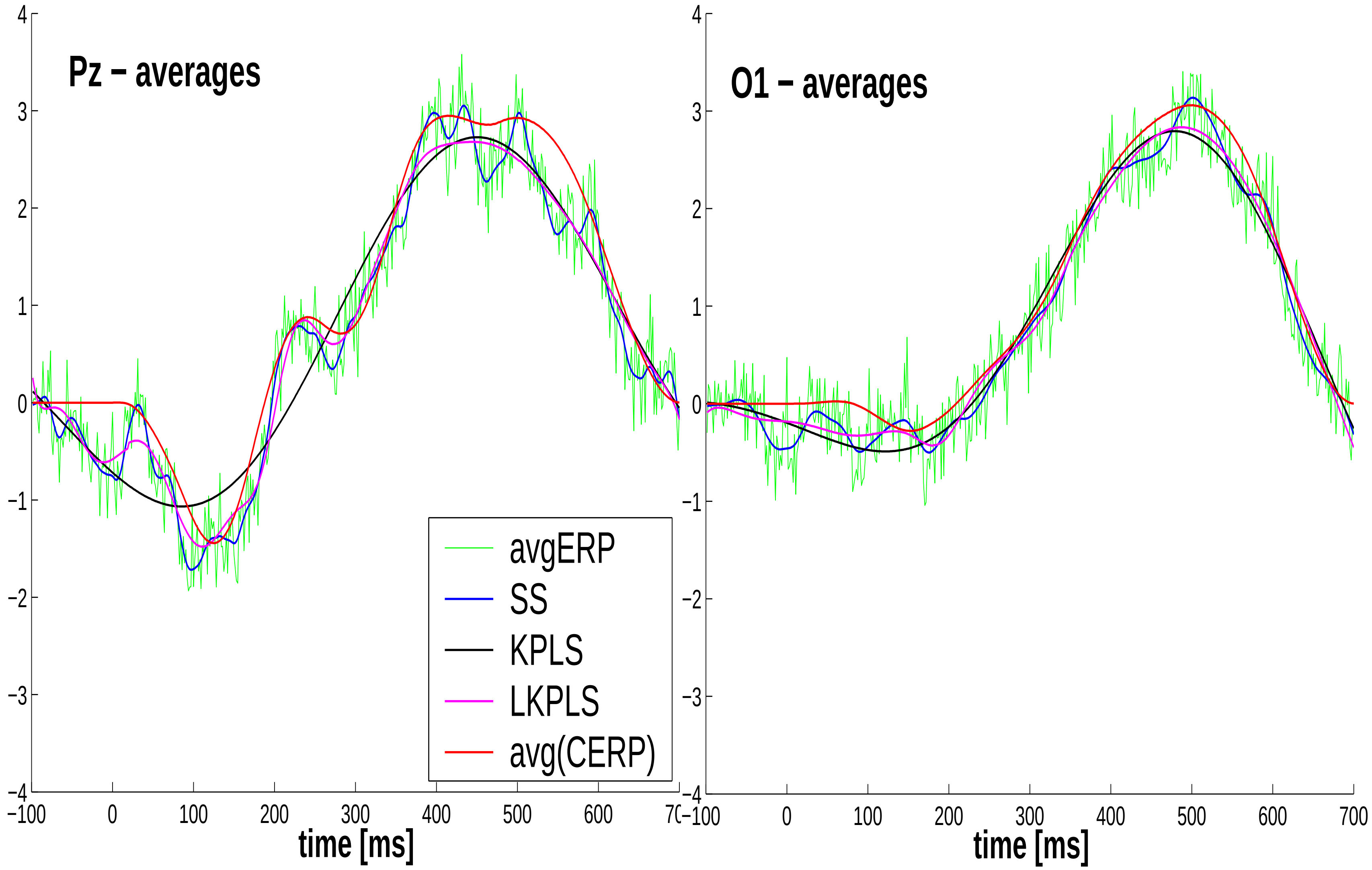
- measured signal$_i$ $=$ ERP$_i$ $+$ (on-going EEG $+$ measur. noise)$_i$
- We compute cov(measured signal$_i$ - avg(measured signal))



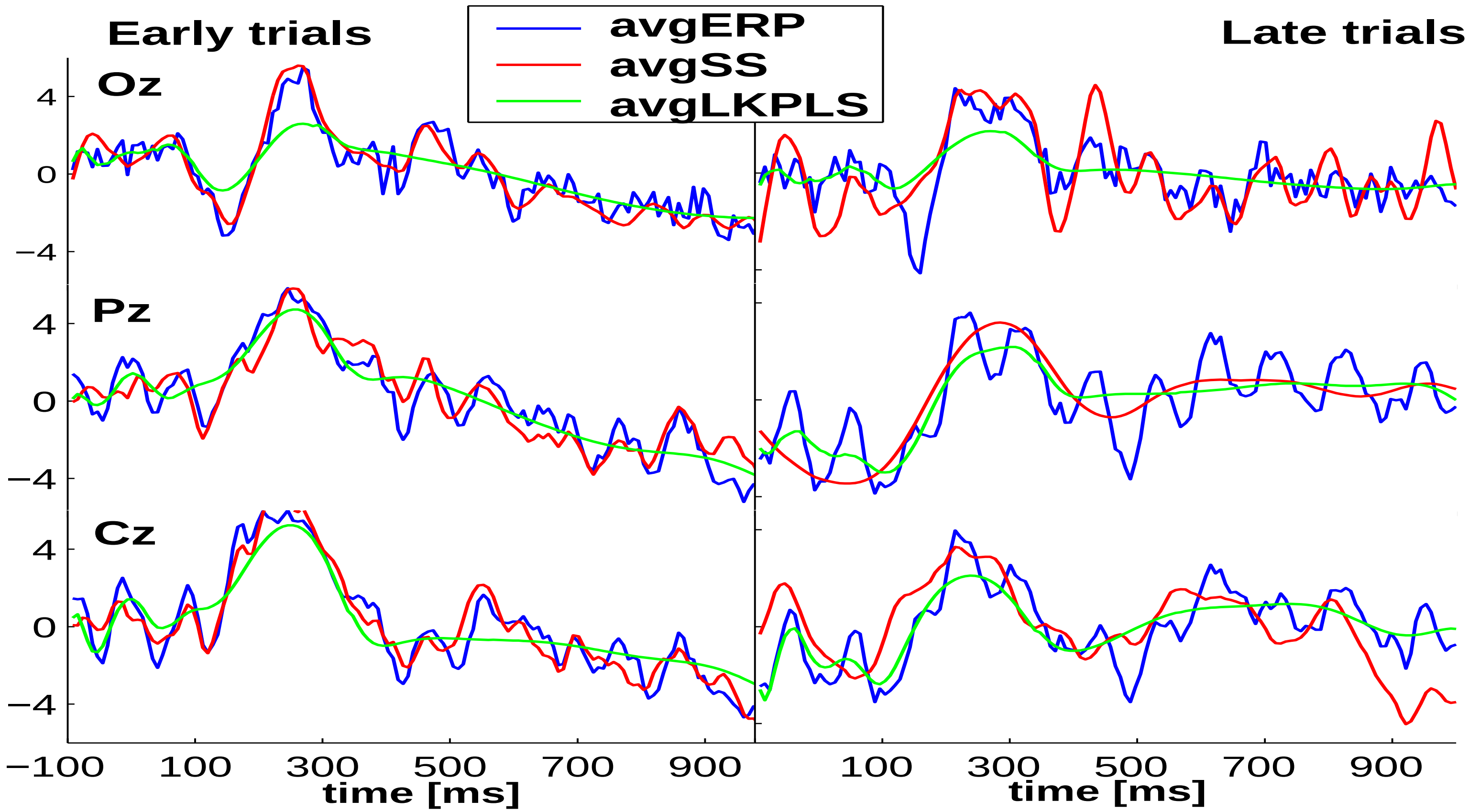
Results on noisy event related potentials (ERPs)–20 different trials were used. Averaged SNR over the trials and electrodes was equal to 1.3dB (min=-7.1dB, max=6.4dB) and 512 samples were used. NRMSE - normalized root mean squared error; SRC - Spearman's rank correlation coefficient.



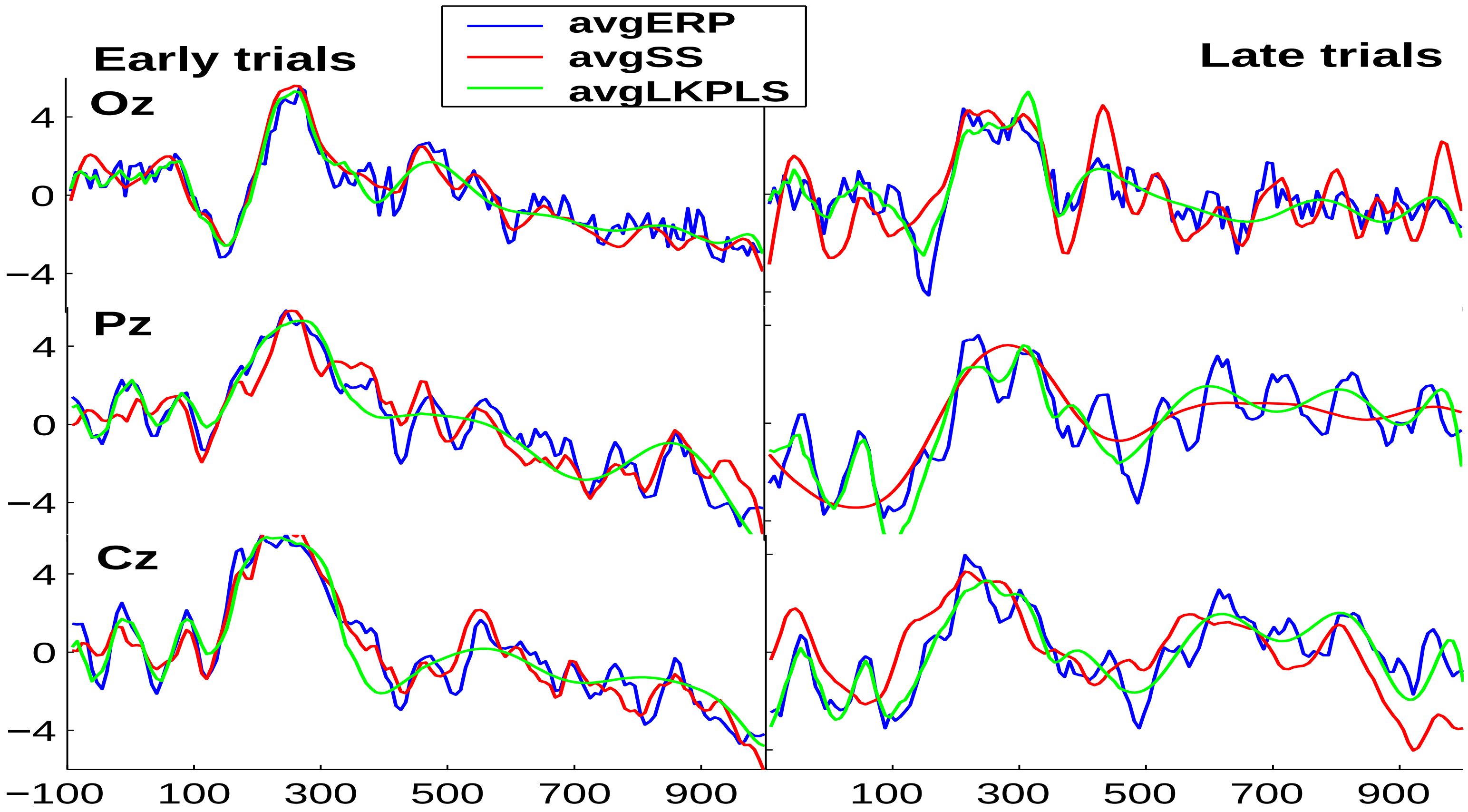
EEG
avgERP
SS
KPLS
LKPLS
ERP
Cz
25% percentile SRC/SS
median SRC/SS
75% percentile SRC/SS
time [ms]
Cz – averages
avgERP
SS
KPLS
LKPLS
avg(CERP)
time [ms]


Pz – averages
O1 – averages
avgERP
SS
KPLS
LKPLS
avg(CERP)
time [ms]
time [ms]
4
3
2
1
0
-1
-2
-3
-4
-100
0
100
200
300
400
500
600
700

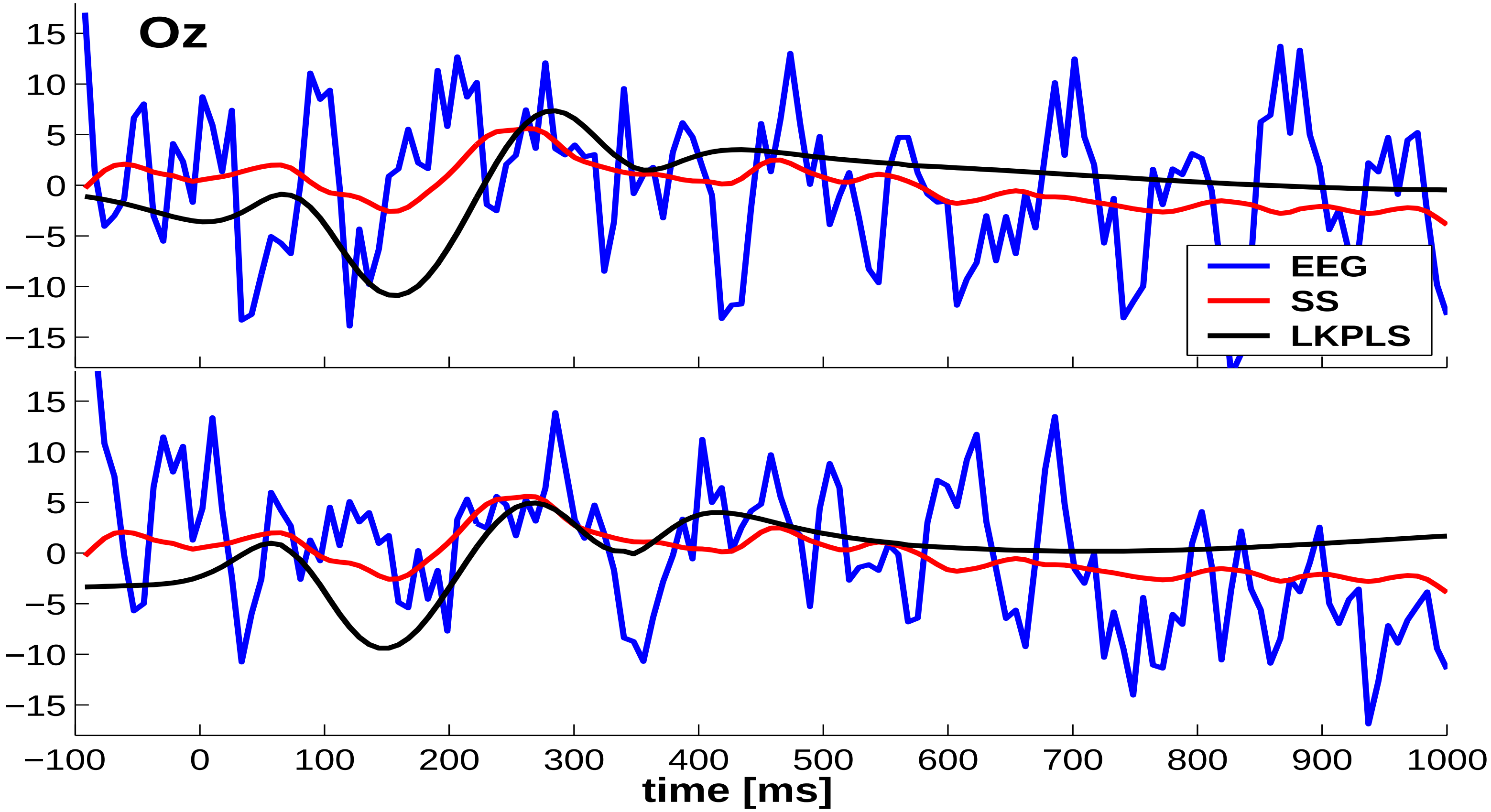
Results on ERPs recorded on a cognitive fatigue experiment



Results on ERPs recorded on a cognitive fatigue experiment



Sample of two ERPs trials recorded on a cognitive fatigue experiment



Conclusions

- PLS Regression - valuable method for data with strong latent structure
- PLS discrimination - useful method for dimensionality reduction, visualization
- PLS - code is simple - do no forget to try it when you look at new data ;-)

References

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

[2] T. De Bie, N. Cristianini, and R. Rosipal. *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*, chapter Eigenproblems in Pattern Recognition. Springer Verlag (in print), 2005.

[3] L. Breiman and J. H. Friedman. Predicting Multivariate Responses in Multiple Regression. *Journal of the Royal Statistical Society: Series B*, 59(1):3–54, 1997.

[4] N.A. Butler and M.C. Denham. The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society: Series B*, 62:585–593, 2000.

[5] S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993.

[6] 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.

[7] I.E. Frank and J.H. Friedman. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, 35(2):109–147, 1993.

[8] P.H. Garthwaite. An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*, 89(425):122–127, 1994.

[9] I.S. Helland. On structure of partial least squares regression. *Communications in Statistics – Elements of Simulation and Computation*, 17:581–607, 1988.

[10] I.S. Helland. Some theoretical aspect of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58:97–107, 2001.

[11] A. Höskuldsson. PLS Regression Methods. *Journal of Chemometrics*, 2:211–228, 1988.

[12] O.C. Lingjaerde and N. Christophersen. Shrinkage Structure of Partial Least Squares. *Scandinavian Journal of Statistics*, 27(3):459–473, 2000.

[13] 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.

[14] R. Manne. Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration. *Chemometrics and Intelligent Laboratory Systems*, 2:187–197, 1987.

[15] A.R. McIntosh, F.L. Bookstein, J.V. Haxby, and C.L. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3:143–157, 1996.

[16] N. Krämer. On the shrinkage behavior of Partial Least Squares Regression. Technical report, Technical University of Berlin, `http://stat.cs.tu-berlin.de/ nkraemer` , 2005.

[17] F.A. Nielsen, L.K. Hansen, and S.C. Strother. Canonical Ridge Analysis with Ridge Parameter Optimization. *NeuroImage*, 7(4):S758, 1998.

[18] 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.

[19] 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.

[20] 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(5):477–491, 1989.

[21] M. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, series B*, 36:111–147, 1974.

[22] 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, series B*, 52(2):237–269, 1990.

[23] H. D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4:147–166, 1976.

[24] 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.

[25] H. Wold. Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiah, editor, *Multivariate Analysis*, pages 391–420. Academic Press, New York, 1966.

[26] H. Wold. Soft Modeling by Latent Variables: The Nonlinear Iterative Partial Least Squares (NIPALS) Approach. In J. Gani, editor, *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett on the occasion of his sixty-fifth birthday*, pages 117–142. Academic Press, London, 1975.

[27] 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.

[28] K.J. Worsley. An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5(4):254–258, 1997.