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:
  \[ X \ (n_{\text{objects}} \times N_{\text{variables}}) \]
  \[ Y \ (n_{\text{objects}} \times M_{\text{responses}}) \]
  - zero-mean

- bilinear decomposition:
  \[ X = TP^T + E \]
  \[ Y = UQ^T + F \]

*where:*
  \( T, U \) matrix of score vectors (LV, components)
  \( P, Q \) matrix of loadings
  \( E, F \) matrix of residuals (errors)
• PLS - bilinear decomposition of $\mathbf{X}$ and $\mathbf{Y}$ maximizing covariance between score vectors $\mathbf{t} = \mathbf{Xw}$ and $\mathbf{u} = \mathbf{Yc}$

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

$$= \text{var}(\mathbf{Xw})[\text{corr}(\mathbf{Xw}, \mathbf{Yc})]^2 \text{var}(\mathbf{Yc})$$

$$= [\text{cov}(\mathbf{t}, \mathbf{u})]^2$$

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

1) $\mathbf{w} = \mathbf{X}^T \mathbf{u} / (\mathbf{u}^T \mathbf{u})$

2) $\|\mathbf{w}\| \to 1$

3) $\mathbf{t} = \mathbf{Xw}$

4) $\mathbf{c} = \mathbf{Y}^T \mathbf{t} / (\mathbf{t}^T \mathbf{t})$

5) $\|\mathbf{c}\| \to 1$

6) $\mathbf{u} = \mathbf{Yc}$

7) go to 1)

• $\mathbf{p} = \mathbf{X}^T \mathbf{t} / (\mathbf{t}^T \mathbf{t})$ ; $\mathbf{q} = \mathbf{Y}^T \mathbf{u} / (\mathbf{u}^T \mathbf{u})$
Overview and Some Aspects of PLS

Bohinj, February 2005
• instead of NIPALS we can solve an eigenproblem:

\[ w \propto X^T u \propto X^T Yc \propto X^T Y Y^T t \propto X^T Y Y^T Xw \]

\[ X^T Y Y^T Xw = \lambda w \quad \text{or} \quad XX^T Y Y^T t = \lambda t \]
\[ t = Xw \quad \text{or} \quad u = Y Y^T t \]

• sequential extraction of \( \{t_i\}_{i=1}^m \)

\[ X_0 = X \]
\[ t_i = X_{i-1} w_i , \quad X_i = X_{i-1} - t_i p_i^T = X - \sum_{j=1}^i t_j p_j^T \]

• deflation schemes define different forms of PLS
Forms of Partial Least Squares

- **PLS1, PLS2**: rank-one approximation of $X, Y$ with a score vector $t$ and vector of loadings $p, q$
  - $X \rightarrow X - tp^T$; $Y \rightarrow Y - tc^T$
  - mutually orthogonal score vectors $t_i$, $i = 1, \ldots, m$
  - $1st \text{ SV}_{i+1} \geq 2nd \text{ SV}_i$ \rightarrow select one score vector at a time

- **PLS Mode A**: rank-one approximation of $X, Y$ with score vectors $t, u$ and vector of loadings $p, q$
  - $X \rightarrow X - tp^T$; $Y \rightarrow Y - uq^T$
  - mutually orthogonal score vectors $t_i, u_i$, $i = 1, \ldots, m$
• **PLS-SB**: SVD of $Y^T X = A\Sigma B^T$
  - $Y^T X \rightarrow Y^T X - \sigma a b^T$
  - mutually orthogonal weight vectors $a_i, b_i$
  - generally not orthogonal score vectors $c_i = Xa_i, d_i = Yb_i$

• **SIMPLS** : (de Jong’93)
  - avoids deflation of $X$; i.e. finds weight vectors $\tilde{w}_i$
    such that $\tilde{T} = X_0 \tilde{W}$
  - SVD of $X_0^T Y_0$ + constraint of mutually orthogonal $\tilde{t}_i$
  - sequence of SVD problems $\tilde{P}_i^\perp X_0^T Y_0$
    $\tilde{P}_i^\perp$ an orthogonal projector onto $\tilde{P}_i = [\tilde{p}_1, \ldots, \tilde{p}_i]$
    where $\tilde{p}_i = X_0^T \tilde{t}_i / (\tilde{t}_i^T \tilde{t}_i)$ 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 \( \Leftrightarrow \) CR

- PLS:
  \[
  \max_{|r|=|s|=1} [\text{cov}(\mathbf{X}_r, \mathbf{Y}_s)]^2 = \max_{|r|=|s|=1} \frac{\text{var}(\mathbf{X}_r)[\text{corr}(\mathbf{X}_r, \mathbf{Y}_s)]^2 \text{var}(\mathbf{Y}_s)}{\text{var}(\mathbf{X}_r)}
  \]

- CCA:
  \[
  \max_{|r|=|s|=1} [\text{corr}(\mathbf{X}_r, \mathbf{Y}_s)]^2
  \]

- PCA:
  \[
  \max_{|r|=1} [\text{var}(\mathbf{X}_r)]
  \]
Canonical Ridge Analysis - CCA $\iff$ PLS

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

- 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: $\gamma_X \in (0, 1), Y \in \mathcal{R}$
PLS Regression (PLS1, PLS2)

• assume: (i) $T$ are good predictors of $Y$
  (ii) the *inner loop* relation $U = T + H$; i.e. $Y$ is a linear function of $T$
  $H$ matrix of residuals (errors)

• linear PLS regression model:
  $Y = TC^T + F^* = XB + F^*$, $F^*$ matrix of residuals (errors)

• $T = XW^* = XW(P^TW)^{-1}$

• $\hat{Y} = XW(P^TW)^{-1}C^T = XB$
\[ \hat{Y} = XW(P^T W)^{-1}C^T = XB \]

- using the existing relations among \( t, u, c, w \):
  \[ B = X^T U(T^T XX^T U)^{-1}T^T Y \]

- train data:
  \[ \hat{Y} = XX^T U(T^T XX^T U)^{-1}T^T Y = TT^T Y = TC^T \]
  single output: \( \hat{y}(x) = c_1 t_1(x) + c_2 t_2(x) + \ldots + c_m t_m(x) \)

- test data:
  \[ \hat{Y}_t = X_t X^T U(T^T XX^T U)^{-1}T^T Y = T_t C^T \]
Overview and Some Aspects of PLS

Bohinj, February 2005

**PLS1 ⇨ Lanczos Method**

- \( b_{PLS}^{(m)} = R^{(m)}[(R^{(m)})^T X^T X R^{(m)}]^{-1}(R^{(m)})^T X^T y \)

- \( R^{(m)} \) - a matrix with orthonormal columns spanning Krylov space \( K^{(m)} = \text{span}\{X^T y, (X^T X)X^T y, \ldots, (X^T X)^{m-1} X^T y\} \)
  \( W^{(m)} = [w_1, w_2, \ldots, w_m] \) is such a candidate

- \( Z^{(m)} = (R^{(m)})^T X^T X R^{(m)} \) is a tridiagonal matrix

- Lanczos method approximate extremal eigenvalues of \( X^T X \) by constructing a sequence of \( Z^{(m)} \); columns of \( R^{(m)} \) are given by a Gram-Schimdt orthonormalization of the first \( m \) columns of \( K^{(m)} \)
**Conjugate Gradients (CG)**

- CG - solves a system of linear equations $A\mathbf{f} = \mathbf{g}$ by minimization of the quadratic form $\frac{1}{2}\mathbf{f}^T A\mathbf{f} - \mathbf{g}^T \mathbf{f}$ ($A$ positive semidefinite)

- for any $\mathbf{f}_0$, the sequence $\mathbf{f}_j$, iterates to the solution $\mathbf{f} = A^{-1}\mathbf{g}$ in $p = \text{rank}(A)$ steps

- the connection between CG and Lanczos method known (Hestens & Stiefel'52; Lanczos'50)

- if $A = \mathbf{X}^T\mathbf{X}; \quad \mathbf{g} = \mathbf{X}^T\mathbf{y} \quad \& \quad \mathbf{f}_0 = 0$ then $\mathbf{b}_{PLS}^{(m)} \Leftrightarrow \mathbf{f}_m$
Kernel PLS Regression

- linear PLS regression in a feature space \( \mathcal{F} \)

- kernel trick: \( K = \Phi \Phi^T \)
  
  where \( \Phi \) is the \((n \times L)\) matrix of the mapped input data:
  
  \( \Phi : x \rightarrow \Phi(x) \in \mathcal{F} \)

- nonlinear kernel-based PLS:
  
  \[
  XX^T YY^T t = \lambda t \Rightarrow \quad KYY^T t = \lambda t
  
  u = YY^T t
  
  or
  
  iterative kernel-based NIPALS algorithm
$g(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$, \( n=100 \)
local kernel PLS (2 x 4 comps.), MSE=0.0019 – red
kernel PLS (8 comps.), MSE=0.0043 – green
"The Peculiar Shrinkage Properties" of PLS1

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

• assume: \( y = Xb + \epsilon \)

\( y \) an \((n \times 1)\) response vector
\( X \) an \((n \times N)\) design matrix
\( b \) an unknown \((N \times 1)\) parameter vector
\( \epsilon \) an \((n \times 1)\) vector of noise, iid elements \( \sim \mathcal{N}(0, \sigma^2) \)
\( y, X \) centered, i.e. \( \mathbf{1}_n^T Y = 0 \) and \( \mathbf{1}_n^T X = \mathbf{0}_N \),
\( \text{rank}(X) = p \leq \min(n - 1, N) \)
\( \text{svd}(X) = UDV^T \); \( \delta_i \) - singular values
\( X^T X = V \Lambda V^T = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \), \( \lambda_i = \delta_i^2 \)
Ordinary Least Squares (OLS)

- $\min_b \|y - Xb\|_2 \Rightarrow \hat{b}_{OLS} = (X^T X)^{-1} X^T Y = V \Lambda^{-1/2} U^T Y$
  \[ \hat{b}_{OLS} = \sum_{i=1}^{p} \lambda_i^{-1/2} (u_i^T y) v_i = \sum_{i=1}^{p} \hat{b}_i \]

- $\hat{b}_{OLS}$ belongs to the class of linear estimators $\hat{z} = Ly$
  \[ E(\hat{z}) = LXz \]
  \[ \text{var}(\hat{z}) = \sigma^2 \text{trace}(LL^T) \]

- $E(\hat{b}_{OLS}) = b$
  \[ \text{var}(\hat{b}_{OLS}) = E[(\hat{b}_{OLS} - b)^T (\hat{b}_{OLS} - b)] = \sigma^2 \text{trace}(X^T X)^{-1} = \]
  \[ = \sigma^2 \sum_{i=1}^{m} \frac{1}{\lambda_i} \]

- $MSE(\hat{z}) = (E(\hat{z}) - z)^T (E(\hat{z}) - z) + E[(\hat{z} - E(\hat{z}))^T (\hat{z} - E(\hat{z}))]$
  \[ \equiv \text{bias}^2(\hat{z}) + \text{var}(\hat{z}) \]

- if $\|\hat{z}_1\|_2 \leq \|\hat{z}_2\|_2 \Rightarrow \text{var}(\hat{z}_1) \leq \text{var}(\hat{z}_2)$
\textbf{Shrinkage Estimators}

\begin{itemize}
  \item $\hat{b}_{shr} = \sum_{i=1}^{p} f(\lambda_i)\lambda_i^{-1/2}(u_i^T y)v_i = \sum_{i=1}^{p} f(\lambda_i)\hat{b}_i$
  \hspace{1cm} $\hat{b}_i$ – the component of $\hat{b}_{OLS}$ along $v_i$

  \item linear shrinkage estimators

  $MSE(\hat{b}_{shr}) = \sum_{i=1}^{p} (f(\lambda_i) - 1)^2(v_i^T b)^2 + \sigma^2 \sum_{i=1}^{p} f(\lambda_i)^2 / \lambda_i$
\end{itemize}
(Generalized) Ridge Regression

\[ f(\lambda_i) = \frac{\lambda_i}{\lambda_i + \gamma_i}, \quad \gamma_i \text{ - regularization term along } v_i \]

Principal Components Regression (PCR)

\[ f(\lambda_i) = \begin{cases} 
1 & : \text{principal component along } v_i \text{ included} \\
0 & : \text{otherwise}
\end{cases} \]
PLS Regression (PLS1)

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

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

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

- $R^{(m)}$ - a matrix with orthonormal columns spanning Krylov space $K^{(m)} = \text{span}\{X^T y, (X^T X) X^T y, \ldots, (X^T X)^{m-1} X^T y\}$

$W^{(m)} = [w_1, w_2, \ldots, w_m]$ is such a candidate
Fundamental Properties of $f^{(m)}(\lambda_i)$

- $f^{(m)}(\lambda_i)$ depends non-linearly on $y$
- $f^{(m)}(\lambda_i) > 1$ 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 $X^TX$)
  (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 $f^{(m)}(\lambda_i) \geq 1$
Overview and Some Aspects of PLS

Bohinj, February 2005

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 $m=11$
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 ⇒ 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

• \( t = Xw \); explained variance (fit) associated with \( t \) is
  \[ r^2 = \frac{(y^T t)^2}{t^T t} \]

• let \( X = [X_1, X_2] \) and weight vector \( w = [w_1, w_2] \):
  \[ (y^T t)^2 = ((y^T X_1 w_1) + (y^T X_2 w_2)) \]
  \[ t^T t = w_1^T X_1^T X_1 w_1 + 2w_2^T X_2^T X_1 w_1 + w_2^T X_2^T X_2 w_2 \]

• problem: large \( (w_2^T X_2^T X_2 w_2) \) can spoil good fit given by large \( (y^T X_1 w_1) \); e.g large amount of small components in \( w \)

• (i) compute \( w \) using \( X \)
  (ii) sort \( x_i \) using \( \text{abs}(w) \)
  (iii) compute \( r^2 \) and/or cross-validate sub-models
  (iv) compute new PLS model \( (w, t, \ldots) \) using selected \( x_i \)
**PLS Discrimination/Classification**

\[ Y = \begin{pmatrix}
  1_{n_1} & 0_{n_1} & \ldots & 0_{n_1} \\
  0_{n_2} & 1_{n_2} & \ldots & 0_{n_2} \\
  \vdots & \vdots & \ddots & \vdots \\
  0_{n_g} & 0_{n_g} & \ldots & 0_{n_g}
\end{pmatrix} \]

**Orthonormalized PLS**

\[ \tilde{Y} = Y(Y^TY)^{-1/2} \]

\[ \tilde{Y}^T\tilde{Y} = I \]
Orthonormalized PLS vs. CCA, Fisher’s LDA

- orthonormalized PLS
  \[
  \max_{|r|=|s|=1} [\text{cov}(X_r, \tilde{Y}_s)]^2 = \text{var}(X_w)[\text{corr}(X_w, \tilde{Y}_c)]^2 \\
  X^T\tilde{Y}\tilde{Y}^TX_w = \lambda w
  \]
  \[
  X^TY(Y^TY)^{-1}Y^TX_w = \lambda w \\
  Hw = \lambda w
  \]

- CCA, Fisher’s LDA
  \[
  \max_{|r|=|s|=1} [\text{corr}(X_r, Y_s)]^2 = [\text{corr}(X_a, Y_b)]^2 \\
  (X^TX)^{-1}X^TY(Y^TY)^{-1}Y^TX_a = \lambda a \\
  E^{-1}Ha = \frac{\lambda}{1-\lambda} a
  \]
Canonical Ridge Analysis - CCA $\Rightarrow$ PLS

\[
([1 - \gamma_X]X^TX + \gamma_X I)^{-1}X^TY([1 - \gamma_Y]Y^TY + \gamma_X I)^{-1}Y^TXw = \lambda w
\]

- 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: $\gamma_X \in (0, 1)$, $Y \in \mathcal{R}$
Overview and Some Aspects of PLS

Bohinj, February 2005
Kernel PLS Discrimination

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

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

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

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 ($Y \subseteq \mathcal{R}$)

1. kernel PLS score vectors extraction
   - compute $K$ - centered Gram matrix
   - set $K_{res} = K$, $m$ - the number of score vectors
   - for $i = 1$ to $m$
     - $t_i = K_{res} Y$
     - $\|t_i\| \rightarrow 1$
     - $u_i = Y(Y^Tt_i)$
     - $K_{res} \leftarrow K_{res} - t_i(t_i^T K_{res})$
     - $Y \leftarrow Y - t_i(t_i^TY)$
   - end

2. projection of test samples
   - $T_t = K_t U (T^T K U)^{-1}$ ; ($K_t$ - test set Gram matrix)
Experiments - Classification

- 13 benchmark data sets of two-class classification problem
  [http://www.first.gmd.de/~raetsch](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)

b) First kernel PLS score vector

c) First kernel PCA principal component
Overview and Some Aspects of PLS

First kernel PLS score vector – $t_1$

Second kernel PLS score vector – $t_2$
<table>
<thead>
<tr>
<th>Data Set</th>
<th>KFD</th>
<th>C-SVC</th>
<th>KPLS-SVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana</td>
<td>10.8±0.5</td>
<td>11.5±0.5</td>
<td>10.5±0.4</td>
</tr>
<tr>
<td>B.Cancer</td>
<td>25.8±4.6</td>
<td>26.0±4.7</td>
<td>25.1±4.5*</td>
</tr>
<tr>
<td>Diabetes</td>
<td>23.2±1.6</td>
<td>23.5±1.7</td>
<td>23.0±1.7</td>
</tr>
<tr>
<td>German</td>
<td>23.7±2.2</td>
<td>23.6±2.1</td>
<td>23.5±1.6</td>
</tr>
<tr>
<td>Heart</td>
<td>16.1±3.4</td>
<td>16.0±3.3</td>
<td>16.5±3.6</td>
</tr>
<tr>
<td>Image</td>
<td>4.76±0.58</td>
<td>2.96±0.60</td>
<td>3.03±0.61</td>
</tr>
<tr>
<td>Ringnorm</td>
<td>1.49±0.12</td>
<td>1.66±0.12</td>
<td>1.43±0.10</td>
</tr>
<tr>
<td>F.Solar</td>
<td>33.2±1.7</td>
<td>32.4±1.8</td>
<td>32.4±1.8</td>
</tr>
<tr>
<td>Splice</td>
<td>10.5±0.6</td>
<td>10.9±0.7</td>
<td>10.9±0.8</td>
</tr>
<tr>
<td>Thyroid</td>
<td>4.20±2.07</td>
<td>4.80±2.19</td>
<td>4.39±2.10</td>
</tr>
<tr>
<td>Titanic</td>
<td>23.2±2.06</td>
<td>22.4±1.0</td>
<td>22.4±1.1*</td>
</tr>
<tr>
<td>Twonorm</td>
<td>2.61±0.15</td>
<td>2.96±0.23</td>
<td>2.34±0.11</td>
</tr>
<tr>
<td>Waveform</td>
<td>9.86±0.44</td>
<td>9.88±0.43</td>
<td>9.58±0.36</td>
</tr>
</tbody>
</table>
Vowel sounds data set: 11 classes, 10 predictors

First kernel PLS score vector – $t_1$

Second kernel PLS score vector – $t_2$
Vowel sounds data set: 11 classes, 10 predictors

- Linear kernel
- Gaussian kernel, w=2

First linear PLS score vector - $t_1$
Second linear PLS score vector - $t_2$
Third linear PLS score vector - $t_3$
First kernel PLS score vector - $t_1$
Second kernel PLS score vector - $t_2$
Third kernel PLS score vector - $t_3$
<table>
<thead>
<tr>
<th>Method</th>
<th>Training Error</th>
<th>Testing Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.32</td>
<td>0.56</td>
</tr>
<tr>
<td>SVC (linear) - 1vs1</td>
<td>0.19</td>
<td>0.51</td>
</tr>
<tr>
<td>KPLS-SVC (linear) - 1vs1</td>
<td>0.16</td>
<td>0.47</td>
</tr>
<tr>
<td>FDA/MARS (df=2)</td>
<td>0.02</td>
<td>0.42</td>
</tr>
<tr>
<td>FDA/MARS (df=6,red. dim.)</td>
<td>0.13</td>
<td>0.39</td>
</tr>
<tr>
<td>SVC (gauss) - 1vs1</td>
<td>0.01</td>
<td>0.37</td>
</tr>
<tr>
<td>KPLS-SVC (gauss) - 1vs1</td>
<td>0.01</td>
<td>0.35</td>
</tr>
<tr>
<td>SVC (gauss, w ≤ 5) - 1vs1</td>
<td>0.002</td>
<td>0.29</td>
</tr>
<tr>
<td>KPLS-SVC (gauss, w ≤ 5) - 1vs1</td>
<td>0.002</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Finger movement periods vs. non-movement periods
Overview and Some Aspects of PLS

Bohinj, February 2005

PLS-derived Spatio-temporal Filter - 01/09/2003
Overview and Some Aspects of PLS

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
Problem solving
? 4 – 15 s

Response

Intertrial interval 1 s

Problem solving
? 4 – 15 s

3 + 5 - 7 + 1 ≤ ≥ 2?

9 - 6 + 2 - 4 ≤ ≥ 2?
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

\[ \min_f \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{a}^{b} (f''(x))^2 \, dx \right) \quad \lambda > 0 \Rightarrow \text{natural cubic splines with knots at } x_i ; i = 1, \ldots, n \]

• Complete basis → **shrink** the coefficients toward smoothing

Wavelet Smoothing

• Complete orthonormal basis → **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 1.3 dB (min=−7.1 dB, max=6.4 dB) and 512 samples were used. NRMSE - normalized root mean squared error; SRC - Spearman’s rank correlation coefficient.
Overview and Some Aspects of PLS

Bohinj, February 2005

Cz

EEG   avgERP
SS     KPLS
LKPLS  ERP

25% percentile SRC/SS
median SRC/SS
75% percentile SRC/SS

Cz – averages

avgERP
SS     KPLS
LKPLS  ERP
avg(CERP)
Results on ERPs recorded on a cognitive fatigue experiment

![Graph showing ERP waveforms for different time points and electrodes (Oz, Pz, Cz) with traces labeled avgERP, avgSS, and avgLKPLS for early and late trials. The graph plots time in ms against voltage.](image-url)
Results on ERPs recorded on a cognitive fatigue experiment

- Early trials
  - Oz
  - Pz
  - Cz

- Late trials
  - avgERP
  - avgSS
  - avgLKPLS

Time units: -100 to 900
Sample of two ERPs trials recorded on a cognitive fatigue experiment

Graph showing ERP responses with time as the x-axis and voltage (in microvolts) as the y-axis, labeled with "Oz".
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


