The Degrees of Freedom of Partial Least Squares Regression

Dr. Nicole Krämer

TU München

5th ESSEC-SUPELEC Research Workshop
May 20, 2011
My talk is about …

… *the statistical analysis of Partial Least Squares Regression.*

1. intrinsic complexity of PLS
2. comparison of regression methods
3. model selection
4. variable selection based on confidence intervals
Example: Near Infrared Spectroscopy

Predict

\( Y \) the percentage of water in meat based on
\( X \) its near infra red spectrum.

▶ unknown linear relationship

\[ f(x) = \beta_0 + \langle \beta, x \rangle = \beta_0 + \sum_{j=1}^{p} \beta_j x^{(j)}. \]

▶ We observe

\[ y_i \approx f(x_i), \quad i = 1, \ldots, n. \]

centered data \( X = \begin{pmatrix} x_1^T \\ \vdots \\ x_p^T \end{pmatrix} \in \mathbb{R}^{n \times p} \quad y = (y_1, \ldots, y_n) \in \mathbb{R}^n. \)
Partial Least Squares (PLS) =

1. supervised dimensionality reduction
2. + least squares regression

The PLS components $T$ have maximal covariance to the response variable $y$.

The $m \ll p$ components $T$ are used as new predictor variables in a least-squares fit.
Partial Least Squares (PLS) =

1. supervised dimensionality reduction
2. + least squares regression

PLS components $T$ have maximal covariance to $y$.

**Algorithm (NIPALS)**

\[ X_1 = X. \text{ For } i = 1, \ldots, m \leftarrow \text{model parameter} \]

1. $w_i \propto X_i^\top y$  
   maximize covariance

2. $t_i \propto X_i w_i$  
   latent component

3. $X_{i+1} = X_i - t_i t_i^\top X_i$  
   enforce orthogonality

\[ \hat{\beta}_m = W (T^\top X W)^{-1} T^\top y \]

Return $T = (t_1, \ldots, t_m)$ and $W = (w_1, \ldots, w_m)$. 
Degrees of Freedom (DoF)

1. capture the intrinsic complexity of a regression method.

\[ Y_i = f(x_i) + \varepsilon_i , \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \]

2. are used for model selection.

\[ \text{test error} = \text{training error} + \text{complexity(DoF)} \]

Examples:

- Bayesian Information Criterion

\[ \text{BIC} = \frac{\| y - \hat{y} \|^2}{n} + \frac{\log(n) \text{var}(\varepsilon)}{n} \text{DoF} \]

- Akaike Information Criterion, Minimum Description Length, ...
**Definition: Degrees of Freedom**

Assumption: The regression method is *linear*, i.e.

\[ \hat{y} = H y, \quad \text{with } H \text{ independent of } y. \]

**Degrees of Freedom**

The Degrees of Freedom of a linear fitting method are

\[ \text{DoF} = \text{trace}(H). \]

**Examples**

- Principal Components Regression with \( m \) components

  \[ \text{DoF}(m) = 1 + m \]

- Ridge Regression, Smoothing splines, ...
Recall: PLS is not linear in $y$

$$\hat{y}_m = \bar{y} + T \left( T^\top T \right)^{-1} T^\top y.$$  

$= \hat{H}$ depends on $y$

 Degrees of Freedom $\neq 1 + \text{trace}(\hat{H}) = 1 + m$.

If we ignore the nonlinearity, we obtain

$$\text{DoF}_{\text{naive}}(m) = 1 + m.$$
The generalized Degrees of Freedom of a regression method are

$$\text{DoF} = E_Y \left[ \text{trace} \left( \frac{\partial \hat{y}}{\partial y} \right) \right]$$

This coincides with the previous definition if the method is linear.

**Proposition**

An unbiased estimate of the Degrees of Freedom of PLS is

$$\hat{\text{DoF}}(m) = \text{trace} \left( \frac{\partial \hat{y}_m}{\partial y} \right).$$

We need to compute (the trace of) the first derivative.
Computational Details (K. & Braun, 2007)

\( m = 1 \) We compute the derivative along the lines of the PLS algorithm

1. \( w_1 = X^\top y \)

\[ \partial w_1 = X^\top \]

2. \( t_1 = Xw_1 \)

\[ \partial t_1 = X \partial w_1 = XX^\top \]

3. \( \hat{y}_1 = P_{t_1} y \)

\[
\partial \hat{y}_1 = \frac{1}{\|t_1\|^2} (t_1 y^\top + t_1 y I_n) (I_n - P_{t_1}) \partial t_1 + P_{t_1}
\]

\( m > 1 \) We rearrange the algorithm in terms of projections \( P \) onto vectors \( t_i \).
Define empirical scatter matrices

\[ s = X^\top y \quad \text{and} \quad S = X^\top X \]

- \( w_1, \ldots w_m \) is an orthogonal basis of the Krylov space

\[ \mathcal{K}_m(S, s) = \text{span} \left( s, Ss, \ldots, S^{m-1}s \right) =: K_m \]

PLS computes orthogonal Gram-Schmidt basis of \( s, Ss, \ldots, S^{m-1}s \)

- Minimization property:

\[ \hat{\beta}_m = \arg \min_{\beta \in \mathcal{K}_m(S, s)} \| y - X\beta \|^2 = K_m \left( K_m^\top S K_m \right)^{-1} K_m^\top s \]

\[ \sim \] explicit formula for the trace of \( \partial \hat{y}_m \) (but not for the derivative itself)
Two equivalent algorithms

- implemented in the R-package `plsdof`.

```r
my.pls<-pls.model(X,y,compute.DoF=TRUE,compute.jacobian)
```

- option: runtime scales in \( p \) (number of variables) or in \( n \) (number of observations).

<table>
<thead>
<tr>
<th>( \partial \hat{y}_m ) and ( \partial \hat{\beta}_m = A? )</th>
<th>iterative computation of ( \partial \hat{y}_m )</th>
<th>projection on Krylov subspaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>confidence intervals</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>for ( \hat{\beta}_m? )</td>
<td>( \text{cov} \left( \hat{\beta}_m \right) = \sigma^2 AA^\top )</td>
<td></td>
</tr>
</tbody>
</table>

The details can be found in the paper.
Benchmark data with different correlation structures.

<table>
<thead>
<tr>
<th>data set</th>
<th>description</th>
<th>variables</th>
<th>examples</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>kin (fh)</td>
<td>dynamics of a robot arm</td>
<td>32</td>
<td>8192</td>
<td>low</td>
</tr>
<tr>
<td>boston</td>
<td>census data for housing</td>
<td>13</td>
<td>506</td>
<td>medium</td>
</tr>
<tr>
<td>cookie</td>
<td>near infrared spectra</td>
<td>700</td>
<td>70</td>
<td>high</td>
</tr>
</tbody>
</table>

```r
go <- pls.model(X, y, compute.DoF=TRUE)
```
The lower is the collinearity, the higher are the Degrees of Freedom.

**Theorem**

If the largest eigenvalue $\lambda_{\text{max}}$ of the sample correlation matrix $S$ fulfills

$$2\lambda_{\text{max}} \leq \text{trace}(S)$$

then

$$\widehat{\text{DoF}}(m = 1) \geq 1 + \frac{\text{trace}(S)}{\lambda_{\text{max}}}.$$
Shape of the DoF-Curve

Benchmark data with different correlation structures.

<table>
<thead>
<tr>
<th>data set</th>
<th>description</th>
<th>variables</th>
<th>examples</th>
<th>training examples</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>kin (fh)</td>
<td>dynamics of a robot arm</td>
<td>32</td>
<td>8192</td>
<td>60</td>
<td>low</td>
</tr>
<tr>
<td>boston</td>
<td>census data for housing</td>
<td>13</td>
<td>506</td>
<td>50</td>
<td>medium</td>
</tr>
<tr>
<td>cookie</td>
<td>near infrared spectra</td>
<td>700</td>
<td>70</td>
<td>39</td>
<td>high</td>
</tr>
</tbody>
</table>

```r
pls.object <- pls.model(X, y, compute.DoF=TRUE)
```

The Degrees of Freedom of PLS
Comparison of Regression Methods: \textit{ozone}

L.A. ozone pollution data
- \( p = 12 \) variables
- \( n = 203 \) observations, \( n_{\text{train}} = 50 \) training observations

Comparison of
1. Partial Least Squares (PLS)
2. Principal Components Regression (PCR)
3. Ridge Regression

\[
\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \{ \| y - X\beta \|^2 + \lambda \| \beta \|^2 \}
\]
There is no difference with respect to mean-squared error.

A direct comparison of model parameters (number of components and $\lambda$) is not possible.

Degrees of Freedom enable a fair model comparison between PLS and PCR.
Comparison of Regression Methods: \textit{ozone}

- PLS fits closer than PCR.” (de Jong)
- However, there is no clear difference with respect to the Degrees of Freedom.
- PLS puts focus on more complex models.
Variable Selection for PLS

PLS does not select variables. \( \leadsto \) extensions to sparse PLS

1. thresholding of the weight vectors (Saigo, K., & Tsuda, 2008)
2. sparsity constraints on the weight vectors (Le Cao et al., 2008; Chun & Keles, 2010)
3. shrinkage (Kondylis & Whittaker, 2007)
4. . . .

Classical approaches

5. bootstrapping (R packages pls and ppls)
6. hypothesis testing based on the distribution of \( \hat{\beta} \)
Approximate distribution of $\hat{\beta}$

Recall: Regression coefficients are a non-linear function of $\mathbf{y}$.

- First order Taylor approximation

\[
\hat{\beta} \approx \frac{\partial \hat{\beta}}{\partial \mathbf{y}} \mathbf{y} =: \mathbf{A}
\]

- approximate covariance matrix

\[
\text{cov} (\hat{\beta}) = \sigma^2 \mathbf{A} \mathbf{A}^T
\]

- The noise level can be estimated via

\[
\hat{\sigma}^2 = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n - \text{DoF}}
\]
Confidence Intervals for PLS: ozone and tecator

- implemented in the R-packages plsdof and multcomp

```r
cv.object <- pls.cv(X, y, compute.covariance = TRUE)
my.multcomp <- glht(cv.object, ...)
```

- The function corrects for multiple comparisons.
- The computational cost is high for large number of variables.
Comparison of

1. 10-fold cross-validation \textit{(gold standard)}
   \[\text{cv.object<-pls.cv}(X, y, k=10)\]

2. Bayesian Information Criterion with our DoF-estimate
   \[\text{bic.object<-pls.ic}(X, y, \text{criterion}='\text{'bic'}})\]

3. Bayesian Information Criterion with the naive estimate
   \[\text{DoF}=m+1\]
   \[\text{naive.object<-pls.ic}(X, y, \text{criterion}='\text{'bic'}}', \text{naive=}\text{TRUE})\]

Akaike Information Criterion and Minimum Description Length are also available in the R-package.

data sets: kin (fh), Boston, cookie
10-fold cross-validation

Bayesian Information Criterion with our DoF-estimate

Bayesian Information Criterion with the naive estimate $\text{DoF}=m+1$

1. All three approaches obtain similar accuracy.
2. There is no clear difference between BIC and naive BIC.
Model complexity (selected components)

- 10-fold cross-validation
- Bayesian Information Criterion with our DoF-estimate
- Bayesian Information Criterion with the naive estimate DoF=m+1

1. BIC selects less complex models than naive BIC.
2. There is no clear difference between BIC and CV.

The plots look similar for the selected Degrees of Freedom.
Why not use the naive approach?

- The naive approach selects more components, but the mean-squared error is not higher. 
  \[\sim\] no overfitting?

- The test error curve can be **flat** or **steep** around the optimum.

![Scaled test error vs components](image)

- Depending on the form of the curve, the selection of too complex models does lead to overfitting.

More details in the paper.
Partial Least Squares ...

- typically consumes more than one Degree of Freedom for each component.

→ precise estimate of its intrinsic complexity

\[
\hat{\text{DoF}}(m) = \text{trace} \left( \frac{\partial \hat{y}_m}{\partial y} \right)
\]

Its Degrees of Freedom ...

- allow us to compare different regression methods.

- select less complex models than the naive estimate (when combined with information criteria).

Variables can be selected ...

- by constructing approximate confidence intervals.
References

The Degrees of Freedom of Partial Least Squares Regression
Journal of the American Statistical Association, in press

plsdof: Degrees of Freedom and Confidence Intervals for Partial Least Squares
R package version 0.2-2

Lanczos Approximations for the Speedup of Kernel Partial Least Squares Regression
Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)

Kernelizing PLS, Degrees of Freedom, and Efficient Model Selection
24th International Conference on Machine Learning (ICML)