27411 Chemometrics and Biological Data Analysis

Lecture 6: PLS - Partial Least Squares (Regression)

Per Bruun Brockhoff

DTU Compute, Statistics and Data Analysis
Building 324, Room 220
Danish Technical University
2800 Lyngby – Danmark
e-mail: perbb@dtu.dk

Overview

1. What is PLS?
2. PLS - How to do it! (same as for PCR)

What is PLS?

1. PLS1: Many X-variables, One Y-variable
   - Alternative to PCR
   - Uses Y-information to construct X-components
2. PLS2: Many X-variables, Many Y-variables
   - Relates components of X to components of Y

PLS1 method:

1. Centering and (possibly) scaling.
2. Do p simple regressions: $\boldsymbol{x}_1$ on y, $\boldsymbol{x}_2$ on y, ..., $\boldsymbol{x}_p$ on y.
3. Use these p regression coefficients $(w_1, w_2, \ldots, w_p)$ to define the first PLS-component, $\boldsymbol{t}_1 = w_1\boldsymbol{x}_1 + w_2\boldsymbol{x}_2 + \cdots w_p\boldsymbol{x}_p$
4. Do the simple regression of $\boldsymbol{y}$ on $\boldsymbol{t}_1$
5. Do the simple regressions of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ on $\boldsymbol{t}_1$
6. Repeat 2-5 on residuals for as many components as wanted/needed.
7. Choose number of components by Cross-Validation

PLS1

- PLS is a *canonical covariance* method
- PLS1 finds X-components with maximal Y-covariance:

$$\max_{||\alpha||=1} \text{Cov}^2(\boldsymbol{y}, \boldsymbol{X}\alpha)$$

- Or equivalently:

$$\max_{||\alpha||=1} \text{Corr}^2(\boldsymbol{y}, \boldsymbol{X}\alpha)\text{Var}(\boldsymbol{X}\alpha)$$

- PCR is a *canonical variance* method
- PCR finds X-components with maximal Y-variance:

$$\max_{||\alpha||=1} \text{Var}(\boldsymbol{X}\alpha)$$

PLS versus PCR

- PLS uses the y-information for building components, PCR does not
- PLS and PCR often predicts on a similar level of error
- PLS often does so with fewer components

How to do it?(same as for PCR)

1. Explore data
2. Do modelling (choose number of components, consider variable selection)
3. Validate (residuals, outliers, influence etc)
4. Iterate e.g. on 2. and 3.
5. Interpret, conclude, report.
6. If relevant: predict future values.

Cross Validation ("Full")

- Leave out one of the observations
- Fit a model on the remaining(reduced) data
- Predict the left out observation by the model: $\hat{y}_{i,val}$
- Do this in turn for ALL observations AND calculate the overall performance of the model:

$$\text{RMSEP} = \sqrt{\sum_{i}^{n}(y_i - \hat{y}_{i,val})^2/n}$$

(Root Mean Squared Error of Prediction)

Cross Validation ("Full")

Choose the optimal number of components:

- The one with overall minimal error
- The first local mininum
- In Hastie et al: the smallest number within the uncertainties of the overall minimum one.

Resampling

- Cross-Validation (CV)
- Jackknifing (Leave-on-out CV)
- Bootstrapping
- A good generic approach:
  - Split the data into a TRAINING and a TEST set.
  - Use Cross-validation on the TRAINING data
  - Check the model performance on the TEST-set
  - MAYBE: REPEAT all this many times (Repeated Double Cross Validation)

Cross Validation - principle

- Minimizes the expected prediction error:

$$\text{Squared Prediction error} = \text{Bias}^2 + \text{Variance}$$

- Including "many" PLS-components: LOW bias, but HIGH variance
- Including "few" PLS-components: HIGH bias, but LOW variance
- Choose the best compromise!
- Note: Including ALL components = MLR (when $n > p$)

Validation - exist on different levels

1. Split in 3: Training(50%), Validation(25%) and Test(25%)
   - Requires many observations - Rarely used
2. Split in 2: Calibration/training (67%) and Test(33%) - us CV/bootstrap within the training
   - more commonly used
3. No "fixed split", but repeated splits by CV/bootstrap, and then CV within each training set ("Repeated double CV")
4. No split, but using (one level of) CV/bootstrap.
5. Just fitting on all - and checking the error.

Overview

DTU Compute
Department of Applied Mathematics and Computer Science