Advanced Methods for Supervised Interval Variable Selection

1. Partial Least Squares Regression
2. LAR/LASSO

KENNESAW STATE UNIVERSITY

Advanced Methods for Supervised Interval Variable Selection

1. Partial Least Squares Regression
2. LAR/LASSO

KENNESAW STATE UNIVERSITY

Objectives

- Describe partial least squares regression.
- Explain how to use partial least squares regression in SAS.
- Discuss advantages and disadvantages with this variable selection method.

KENNESAW STATE UNIVERSITY

Partial Least Squares Regression

- Target used or not?
  - Used
- Original or constructed variables as output?
  - Original variables

Partial Least Squares Regression: Main Features

- As in principal component analysis (PCA), the partial least squares (PLS) algorithm extracts components as a linear combination of the original input variables.
- PCA: Constructs components explaining as much of the variation as possible in the input variables.
- PLS: Constructs components explaining as much of the variation as possible of both the target and input variables.

PLS Compared to Ordinary Least Squares

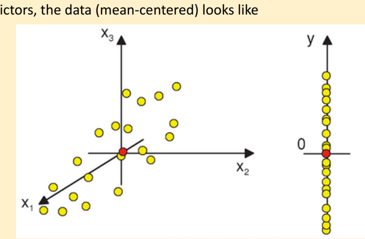
- OLS:
  - The objective is to minimize the prediction error, looking for a linear combination of the input variables, so that as much of the response variation as possible is explained.
- PLS:
  - An additional objective is to account for the variation in the input variables.

PLS Components

- Linear combinations of the input variables are extracted successively.
- Components are also called *factors*, *latent vectors*, *latent variables, PLS components, PLS score*.
- PLS is also used as an abbreviation for *projection to latent structures*.

The Geometry of PLS in the Case of One Response

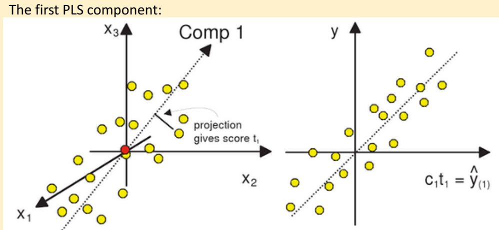
With 3 predictors, the data (mean-centered) looks like



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

The Geometry of PLS in the Case of One Response

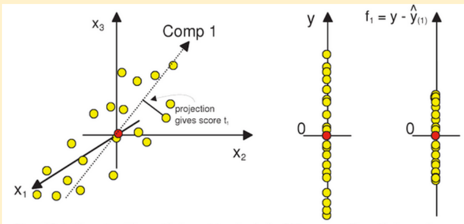
The first PLS component:



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

The Geometry of PLS in the Case of One Response

Residuals after modeling with the first PLS component:

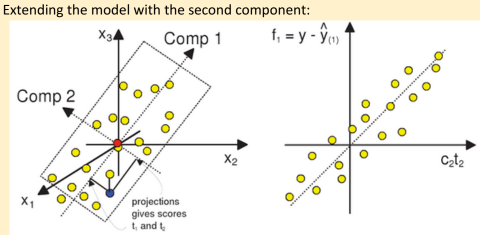


Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

Kennesaw State University

The Geometry of PLS in the Case of One Response

Extending the model with the second component:

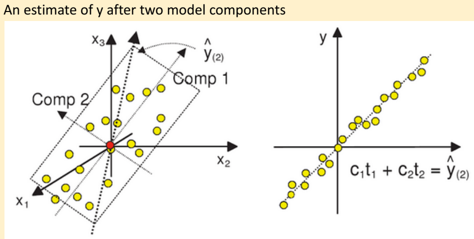


Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

Kennesaw State University

The Geometry of PLS in the Case of One Response

An estimate of y after two model components

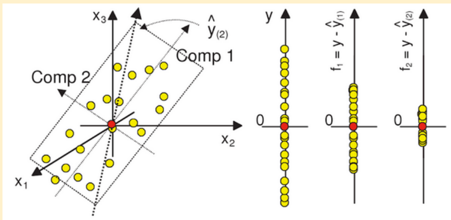


Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

Kennesaw State University

The Geometry of PLS in the Case of One Response

The explanatory power of a PLS model



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

PLS Optimization

(many predictors, **one** response)

- PLS seeks to find linear combinations of the independent variables that summarize the maximum amount of co-variability with the response.
  - These linear combinations are often called *PLS components*, PLS *scores, factors, latent variables, etc*.
  - A PLS *direction* is a vector that points in the direction of maximum co-variance.

PLS Optimization

(many predictors, **one** response)

- PLS is inherently an optimization problem, which is subject to two constraints
  1. The PLS directions have unit length
  2. Either
     a. Successively derived scores are uncorrelated to previously derived scores, OR
     b. Successively derived directions are orthogonal to previously derived directions

Mathematically Speaking…

- The optimization problem defined by PLS can be solved through the following formulation:

$$\underset{a}{\arg\max}\ \frac{\mathrm{Cov}^2\left(a^T X, y\right)}{a^T a},$$

  subject to constraints 2a. or b.

- Facts…
  - the i^{th} PLS direction, a_i, is the eigenvector corresponding to the i^{th} largest eigenvalue of $Z^T Z$, where Z = cov(X,y).
  - the i^{th} largest eigenvalue is the amount of co-variability summarized by the i^{th} PLS component.
  - $a_i^T X$ are the i^{th} scores (the i^{th} PLS component)

Kennesaw State University

PLS *is* Simultaneous Dimension Reduction and Regression

$$\begin{aligned}
&\underset{a}{\arg\max}\ \frac{\mathrm{Cov}^2\left(a^T X, Y\right)}{a^T a} \\
&= \underset{a}{\arg\max}\ \frac{\mathrm{var}\left(a^T X\right)\mathrm{var}(Y)\mathrm{corr}^2\left(a^T X, Y\right)}{a^T a} \\
&= \mathrm{var}(Y)\underset{a}{\arg\max}\ \frac{\mathrm{var}\left(a^T X\right)\mathrm{corr}^2\left(a^T X, Y\right)}{a^T a} \\
&= \mathrm{var}(\text{response})\underset{a}{\arg\max}\ \frac{\mathrm{var}(\text{scores})\mathrm{corr}^2(\text{scores}, \text{response})}{a^T a}
\end{aligned}$$

Kennesaw State University

PLS *is* Simultaneous Dimension Reduction and Regression

max Var(scores) Corr²(response,scores)

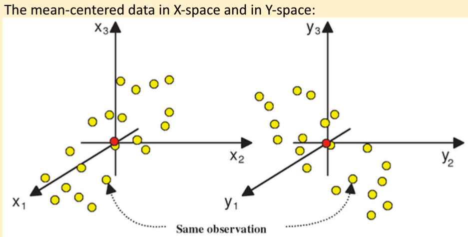
Dimension Reduction (PCA)

Regression

Kennesaw State University

The Geometry of PLS with Multiple Responses

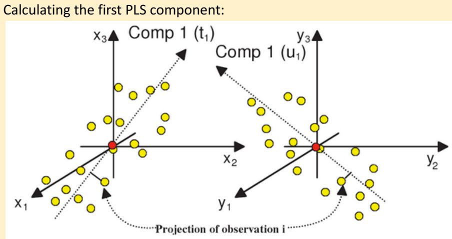
The mean-centered data in X-space and in Y-space:



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

The Geometry of PLS with Multiple Responses

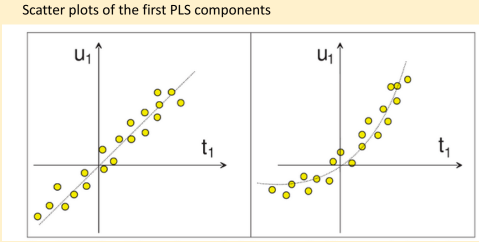
Calculating the first PLS component:



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

The Geometry of PLS with Multiple Responses

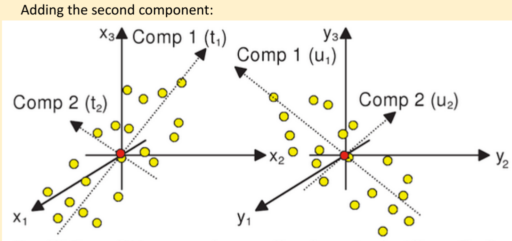
Scatter plots of the first PLS components



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

The Geometry of PLS with Multiple Responses

Adding the second component:

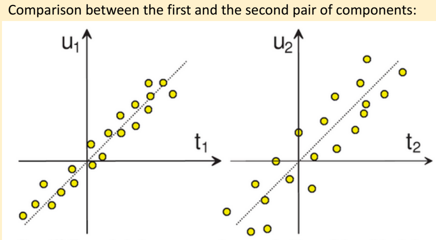


Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

KENNESAW STATE UNIVERSITY

The Geometry of PLS with Multiple Responses

Comparison between the first and the second pair of components:

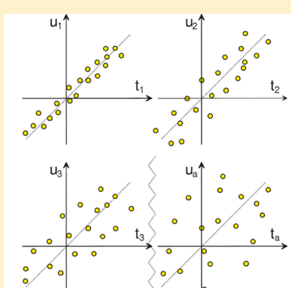


Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

KENNESAW STATE UNIVERSITY

The Geometry of PLS with Multiple Responses

Higher order components:



Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

KENNESAW STATE UNIVERSITY

PLS Optimization (2)

(many predictors, **many** responses)

- PLS seeks to find linear combinations of the independent variables ***and*** a linear combination of the dependent variables that summarize the maximum amount of co-variability between the combinations.
  - These linear combinations are often called *PLS X-space and Y-space components* or *PLS X-space and Y-space scores*.
  - Likewise, X-space and Y-space PLS *directions* point in the direction of maximum co-variance between the spaces.

KENNESAW STATE UNIVERSITY

PLS Optimization (2)

(many predictors, **many** responses)

- PLS is inherently an optimization problem, which is subject to two constraints
  1. The X-space and Y-space PLS directions have unit length
  2. Either
     a. Successively derived scores in each space are uncorrelated to previously derived scores, OR
     b. Successively derived directions in each space are orthogonal to previously derived directions
     - Constraint 2.a. is most commonly implemented

KENNESAW STATE UNIVERSITY

Mathematically Speaking…

- The optimization problem defined by PLS can be solved through the following formulation:

$$\arg\max_{a,b} \frac{\mathrm{Cov}^2\left(a^T X, b^T Y\right)}{\left(a^T a\right)\left(b^T b\right)},$$

$$= \arg\max_{a,b} \frac{\mathrm{var}\left(a^T X\right)\mathrm{var}\left(b^T Y\right)\mathrm{corr}^2\left(a^T X, b^T Y\right)}{\left(a^T a\right)\left(b^T b\right)}$$

subject to constraints 2a. or b.

KENNESAW STATE UNIVERSITY

PLS *is* Simultaneous Dimension Reduction and Regression

max Var(X-scores) Corr^2(X-scores,Y-scores)Var(Y-scores)

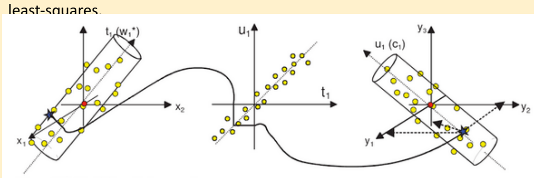
X-space Dimension Reduction (PCA)

Regression

Y-space Dimension Reduction (PCA)

PLS – the "Inner Relation"

The way PLS works *visually* is by "tweeking" the two PCA models (X and Y) until their covariance is optimised. It is this "tweeking" that led to the name *partial* least-squares.



All 3 are solved simultaneously via numerical methods

Reference: *Multi- and Megavariate Data Analysis Basic Principles and Applications*

PLS History

- H. Wold (1966, 1975)
- S. Wold and H. Martens (1983)
- Stone and Brooks (1990)
- Frank and Friedman (1991, 1993)
- Hinkle and Rayens (1994)
- De Jong (1993) – SIMPLS Algorithm

Number of Factors to Extract

- The number of factors that you need depends on the data.
  - Too few factors: Underfitting
  - Too many factors: Overfitting
- Resampling can be used to help determine the number of factors.
- Resampling **the training compounds**

Resampling the Training Component

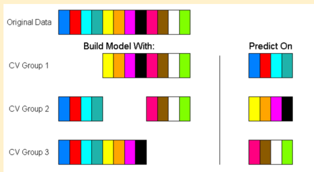
- Resampling only affects the training data
  - The test set is not used in this procedure
- Resampling methods try to "embed variation" in the data to approximate the model's performance on future compounds
- Common resampling methods:
  - K-fold cross validation
  - Leave group out cross validation
  - Bootstrapping

K-fold Cross Validation

- Here, we randomly split the data into *K* blocks of roughly equal size
- We leave out the first block of data (the held-out block) and fit a model on the remained data.
- This model is used to predict the held-out block
- We continue this process until we've predicted all *K* hold-out blocks
- The final performance is based on the hold-out predictions

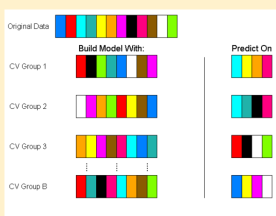
K-fold Cross Validation

- The schematic below shows the process for K = 3 groups.
  - K is usually taken to be 5 or 10
  - leave one out cross-validation has each sample as a block



Leave Group (or p obs.) Out Cross Validation

- A random proportion of data (say 80%) are used to train a model
- The remainder is used to predict performance
- This process is repeated many times and the average performance is used

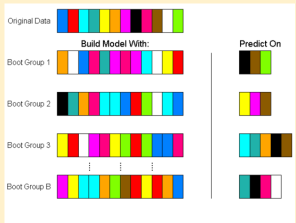


Bootstrapping

- Bootstrapping takes a random sample with replacement
  - the random sample is the same size as the original data set
  - observations may be selected more than once
  - each observation has a 63.2% change of showing up at least once
- Some samples won't be selected
  - these samples will be used to predict performance
- The process is repeated multiple times (say 30)

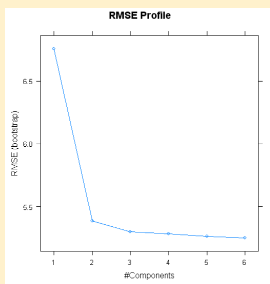
The Bootstrap

- With bootstrapping, the number of held-out samples is random
- Some models, such as random forest, use bootstrapping within the modeling process to reduce over-fitting



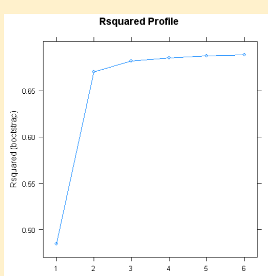
Number of Factors to Extract

- PLS seeks to find latent variables (LVs) that summarize variability and are highly predictive of the response.
- How do we determine the number of LVs to compute?
  - Evaluate RMSPE (or Q^2)
- The optimal number of components is the number of components that minimizes RMSPE



Number of Factors to Extract

- Roughly the same profile is seen when the models are judged on $\mathbf{R^2}$



Number of Factors to Extract

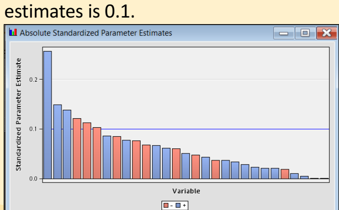
- In EM, the default number of factors to extract is 15.
- The maximum number of factors possible to extract equals the number of input variables (including levels for categorical variables).
- If the maximum number of factors is used, all of the variation in the input variables is explained. The whole variation in the response variable is not explained.

Variable Selection Criterion

- To select which variables should be passed to the successor steps of the Partial Least Squares, two different criteria can be used:
  - Estimated regression coefficients
  - Variable importance for projection (VIP)
- The criteria are selected in the Variable Selection Criterion property. There are four options:
  - Coefficient
  - Variable Importance (Wold. et, al. 1993)
  - Both
  - Either

Estimated Regression Coefficients

- Estimated regression coefficients show the importance for each input variable for the prediction of the target variable.
- The default cutoff for the absolute standardized parameter estimates is 0.1.



Variable Importance for Projection (VIP)

- VIP is a weighted sum of squares of the PLS weights, w*, with the weights calcuated from the amound of Y-variance of each PLS component.
- The VIP score for the jth predictor (in the one-response case)

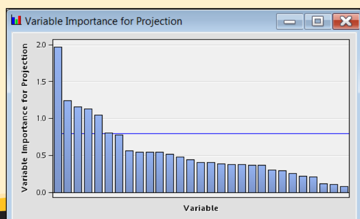
$$VIP_j = \sqrt{\frac{p}{\sum_{m=1}^{M} SS(b_m t_m)} \cdot \sum_{m=1}^{M} w_{mj}^2 SS(b_m t_m)}$$

  - P: # of predictors (x-variables)
  - M: # of retained latent variables
  - W_{mj}: the PLS weight of the j-th variable for the m^{th} latent variable
  - $SS(b_m t_m)$: percent of y explained by the m^{th} latent variable
  - b_m: the regression coeffients on the m^{th} latent variable
  - t_m: the m^{th} latent variable

Kennesaw State University

Variable Importance for Projection (VIP)

- The VIP represents the importance of each input variable for explaining the variation for both target and input variables.
- The default cutoff in EM is 0.8.
- The "greater than 1" rule is also common because the average of squared VIP is 1.



Discussion

- Have you ever performed a PLS regression before?
- What was the business application?
- Why did you do PLS as opposed to using a standard regression model?

Kennesaw State University

Partial Least Squares Regression: Pros

- Often a small number of variables are kept in order to explain a lot of the variation in the data cloud.
- Both the variation in the input variables and the target variable are taken into account – simultaneous dimension reduction and regression.
- Some original variables, not constructed components, are kept as inputs to the next step.

Partial Least Squares Regression: Cons

- How many factors do you extract?
- Factors might not have meaningful interpretation.
- **Cannot do better than ordinary least squares on sample data.**
- Changes to the Variable Selection Criterion property can greatly affect the results (in other words, the final number of inputs selected). (Is this a pro or con?)
- PLS directions will be drawn to independent variables with the most variability (although this will be tempered by the need to also be related to the response)
- Outliers may have significant influence on the directions, resulting scores, and relationship with the response.

Partial Least Squares Regression for Variable Selection

This demonstration illustrates how to use the Partial Least Squares for variable selection.

Reference

– *SAS® Course Materials*

– T. Byrne, E. Johansson, J. Trygg, C. Vikström*, Multi- and Megavariate Data Analysis, L. Eriksson.*

– Svante Wold, Michael Sjöström, Lennart Eriksson, *PLS Regression: A Basic Tool Of Chemometrics.*

– Kee Siong Ng, *A Simple Explanation of Partial Least Squares.*