# 27411 Chemometrics and Biological Data Analysis

## Lecture 4: PCR - Principal Component 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


DTU Compute
Department of Applied Mathematics and Computer Science



## Overview



DTU Compute
Department of Applied Mathematics and Computer Science





## What is PCR? (PCR = PCA + MLR)

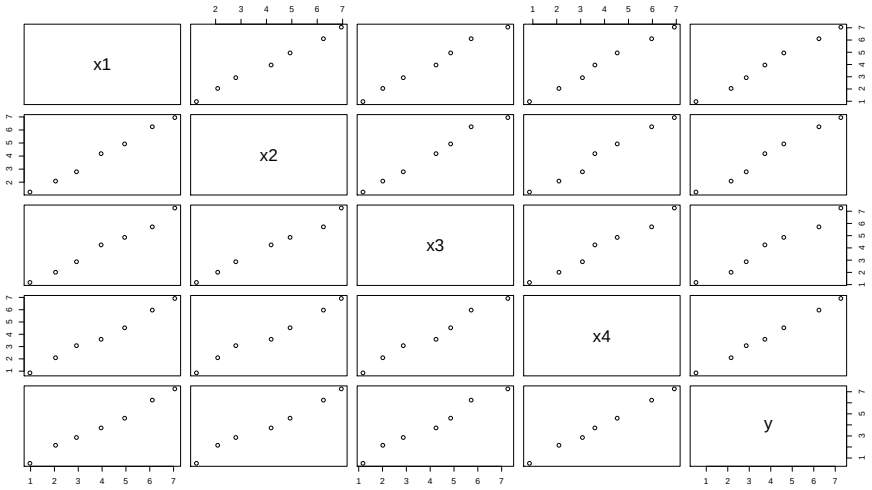
- NOT: *Polymerase Chain Reaction*
- A regression technique to cope with many x-variables
- Situation: Given Y and X-data:
- Do PCA on the X-matrix
  - Defines new variables: the principal components (scores)
- Use some of these new variables in an MLR to model/predict Y
- Y may be univariate OR multivariate: In this course: only UNIVARIATE.

DTU Compute
Department of Applied Mathematics and Computer Science





## Example, Data 1




DTU Compute
Department of Applied Mathematics and Computer Science

## MLR results: $Ey_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_2 x_{i3} + b_p x_{i4}$

| | Estimate | Std. Error | t | P-val |
|---|---|---|---|---|
| Intercept | 0.1589 | 0.1404 | 1.132 | 0.375 |
| x1 | 0.1763 | 0.1347 | 1.309 | 0.321 |
| x2 | -0.2457 | 0.1765 | -1.392 | 0.299 |
| x3 | 0.4068 | 0.2826 | 1.439 | 0.287 |
| x4 | 0.6472 | 0.3168 | 2.043 | 0.178 |

$R^2 = 0.9993, \quad F_{Model} = 675 (\text{p-value} = 0.001480)$





## MLR - remove variables one by one:

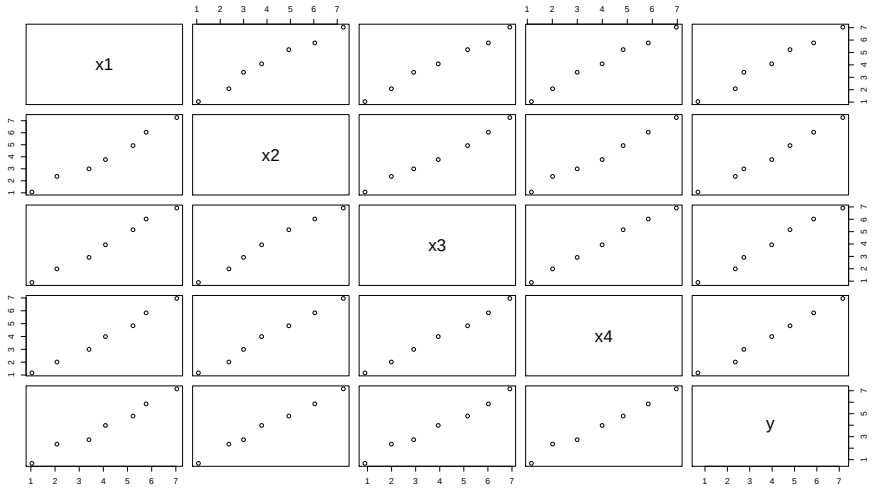
| | Estimate | Std. Error | t | P-val |
|---|---|---|---|---|
| Intercept | -0.02310 | 0.10685 | -0.216 | 0.837 |
| x4 | 1.02855 | 0.02408 | 42.710 | 0.000000133 |

$R^2 = 0.9973, \quad F_{Model} = 1824 (\text{p-value} = 0.000000133)$





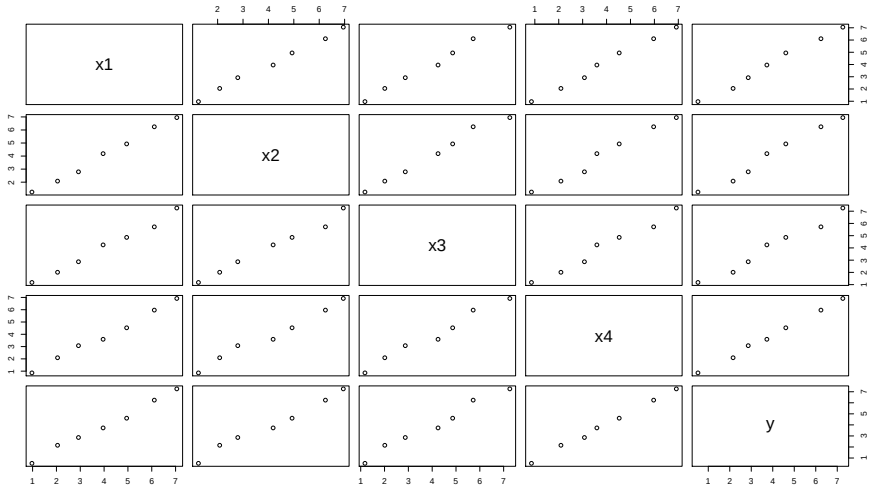
## Example, Data 2








## Example, Data 1



## MLR results, Data 2:

| | Estimate | Std. Error | t | P-val |
|---|---|---|---|---|
| Intercept | -0.1668 | 0.4309 | -0.387 | 0.736 |
| x1 | -0.8141 | 1.6888 | -0.482 | 0.677 |
| x2 | -0.1027 | 0.8635 | -0.119 | 0.916 |
| x3 | 2.0695 | 1.4133 | 1.464 | 0.281 |
| x4 | -0.1354 | 0.6183 | -0.219 | 0.847 |

$R^2 = 0.9877, \quad F_{Model} = 40.26(\text{p-value} = 0.02439)$

DTU Compute
Department of Applied Mathematics and Computer Science





## MLR results, Data 2:

| | Estimate | Std. Error | t | P-val |
|---|---|---|---|---|
| Intercept | -0.14774 | 0.25808 | -0.572 | 0.592 |
| x3 | 1.01508 | 0.05739 | 17.686 | 0.0000106 |

$R^2 = 0.9843, \quad F_{Model} = 312.8(\text{p-value} = 0.0000106)$

DTU Compute
Department of Applied Mathematics and Computer Science





## Use average: $xmn = (x_1 + x_2 + x_3 + x_4)/4$

(For these data: Average $\approx$ 1st Principal Component)

| Data 1 | Estimate | Std. Error | t | P-val |
|---|---|---|---|---|
| Intercept | 0.24606 | 0.11809 | 2.084 | 0.0916 |
| xmn | 0.97818 | 0.02674 | 36.586 | 0.000000287 |

$R^2 = 0.9963, \quad F_{Model} = 1339(\text{p-value} = 0.000000287)$

| Data 2 | Estimate | Std. Error | t | P-val |
|---|---|---|---|---|
| Intercept | -0.17598 | 0.31791 | -0.554 | 0.604 |
| xmn | 1.02418 | 0.07099 | 14.426 | 0.0000289 |

$R^2 = 0.9765, \quad F_{Model} = 208.1(\text{p-value} = 0.0000289)$

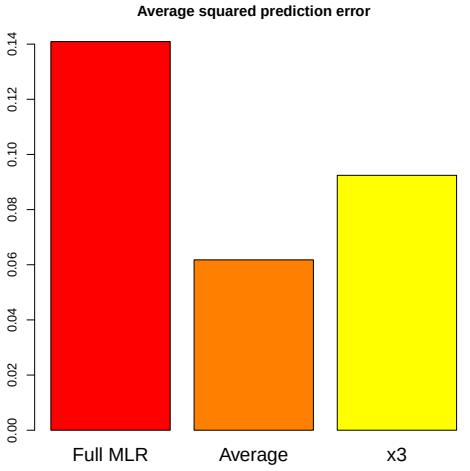
DTU Compute
Department of Applied Mathematics and Computer Science





## Prediction of future values

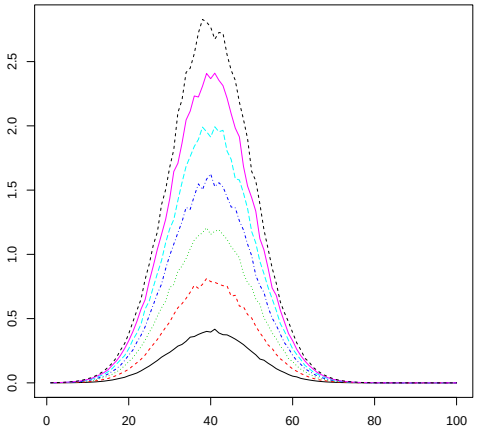
Simulation of 7000 new observations




DTU Compute
Department of Applied Mathematics and Computer Science

## Example: Spectral data

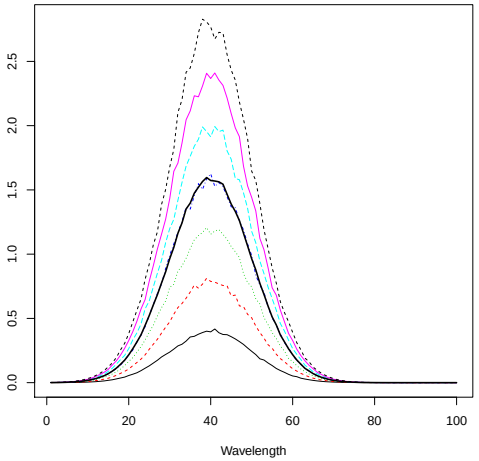
Y-data (PH): $y_1, \ldots, y_7$, X-data: Seven 100-dimensional spectra:







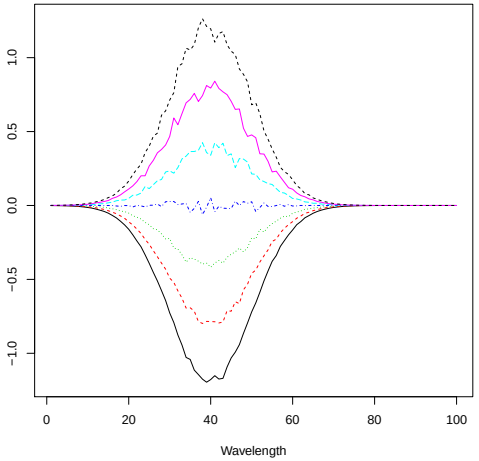

## Spectral data wih mean spectrum







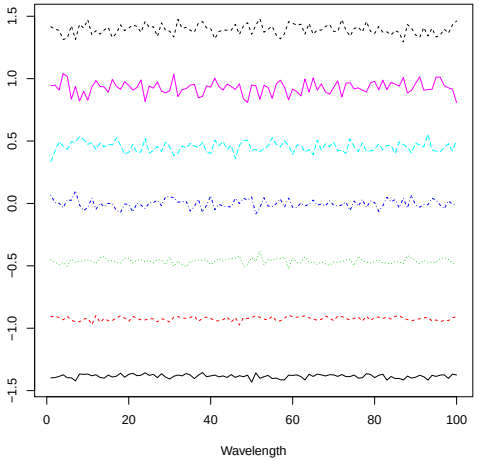

## Mean corrected spectra



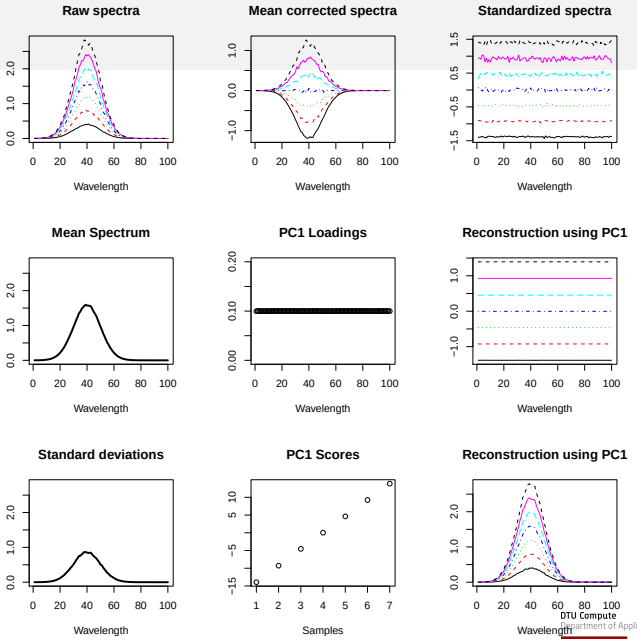





## Standardized AND mean corrected spectra



## PCA:








## PCR: what is it?

- Data Situation:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{np} \end{bmatrix}$$

- Do MLR with $A$ principal components $t_1, \ldots, t_A$ instead of all (or some) of the $x$'s.
- How many components: Determine by Cross-validation!





## How to do it?

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")

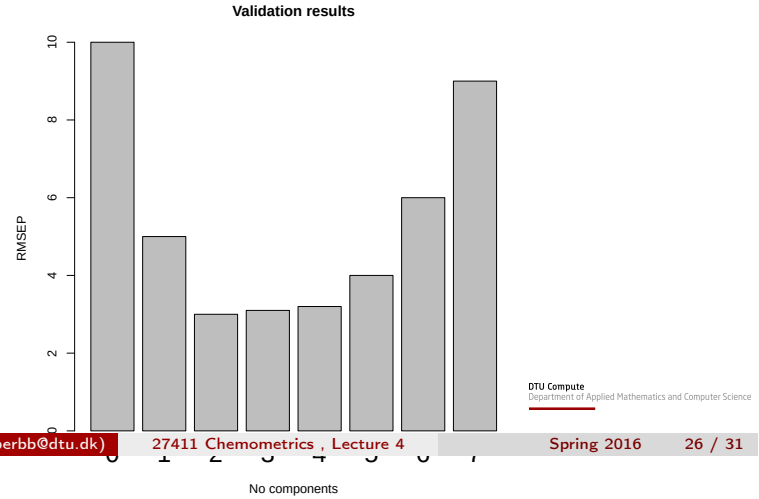
- Leave out one of the observations
- Fit a model on the remaining(reduced) data
- Predict the left out observation by the model: $\hat{y}_{i,val}$
- Do this in turn for ALL observations AND calculate the overall performance of the model:

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

(Root Mean Squared Error of Prediction)

## Cross Validation ("Full")

Finally: Do the cross-validation for ALL choices of number of components $(0, 1, 2, \ldots, \ldots)$ AND plot the performances:








## 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" PC-components: LOW bias, but HIGH variance
- Including "few" PC-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