EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF
SCIENCE
Institute of Physical and Theoretical Chemistry

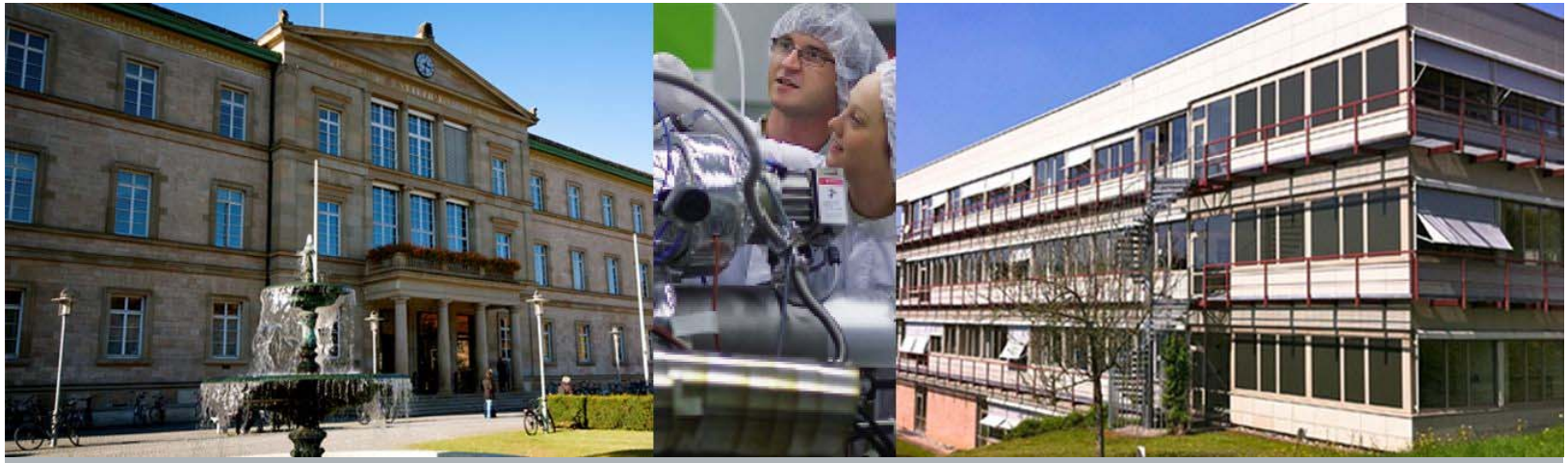


Laboratory Techniques and Medical Device Approvals I

Part 04: PCR

 © 2012 University of Tuebingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

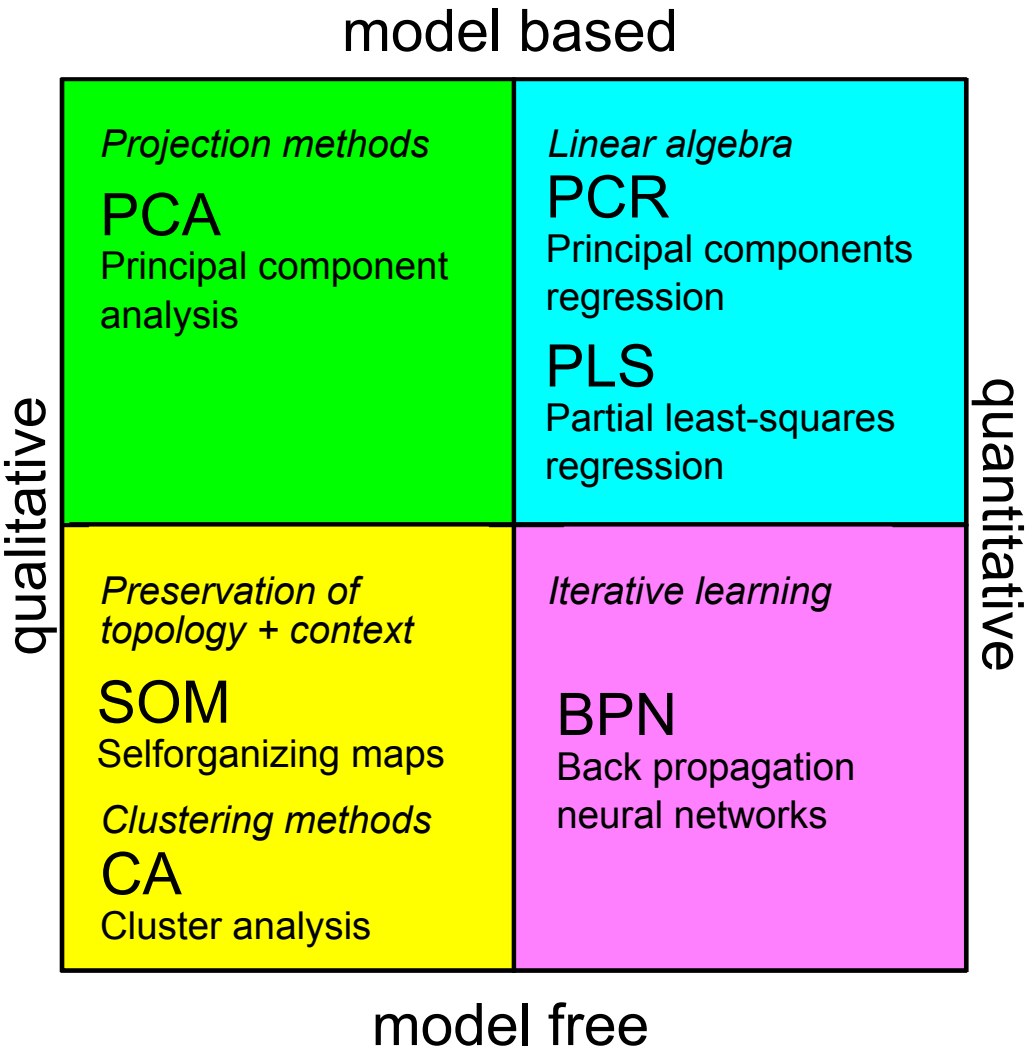


Multivariate Data Analysis

Quantitative Data Analysis:
MLR, PCR

© 2012 University of Tuebingen

Data analysis: Overview over different methods



 © 2012 University of Tuebingen

Regression: Basic considerations

- Besides X information (data matrix, see for PCA) now also Y information (independent variable, e.g. concentration of the measured value)

- Concept of independent test data:

  Data matrix to be split for model building (calibration) and independent test of the model (validation)

- Mulitvariate Linear Regression (MLR)

© 2012 University of Tuebingen

Univariate regression: Example to calculate

With chromatographic methods one should calculate the amount of hydrocarbons. Please use the given table to develop the calibration function.

Mol-percent iso-octane, x_i	Peak area, y_i
0.88	1.363
2.01	2.225
2.70	3.250
3.45	3.788
4.38	5.013

Please calculate:

- *Axis intercept*
- *Slope of the regression line*
- *Standard deviation of beta*

Furthermore use this regression to calculate the amount of iso-octane if the peak area shows 2.650.

The last task is to compute the standard deviation of the value under the following conditions:

- *it was the result of one measurement*
- *it was the mean out of four measurements*

© 2012 University of Tuebingen

Some formulas that might help

Equation for regression line:

$$y = \beta \cdot x + \alpha$$

Axis intercept:

$$\alpha = \bar{y} - \beta\bar{x}$$

Slope of the regression line:

$$\beta = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_i (x_i - \bar{x})^2}$$

Standard deviation of regression:

$$\mathrm{sdv}(x/y) = \sqrt{\frac{\sum_i [y_i - (\beta \cdot x_i + \alpha)]^2}{n-2}}$$

Standard deviation of α:

$$\mathrm{sdv}(\alpha) = \mathrm{sdv}(x/y) \cdot \sqrt{\frac{\sum_i x_i^2}{n \cdot \sum_i (x_i - \bar{x})^2}}$$

Standard deviation of a value:

$$sdv(x_0) = \frac{sdv(x/y)}{\beta} \cdot \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{\beta^2 \cdot \sum_i (x_i - \bar{x})^2}}$$

Standard deviation of β:

$$sdv(\beta) = \frac{sdv(x/y)}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

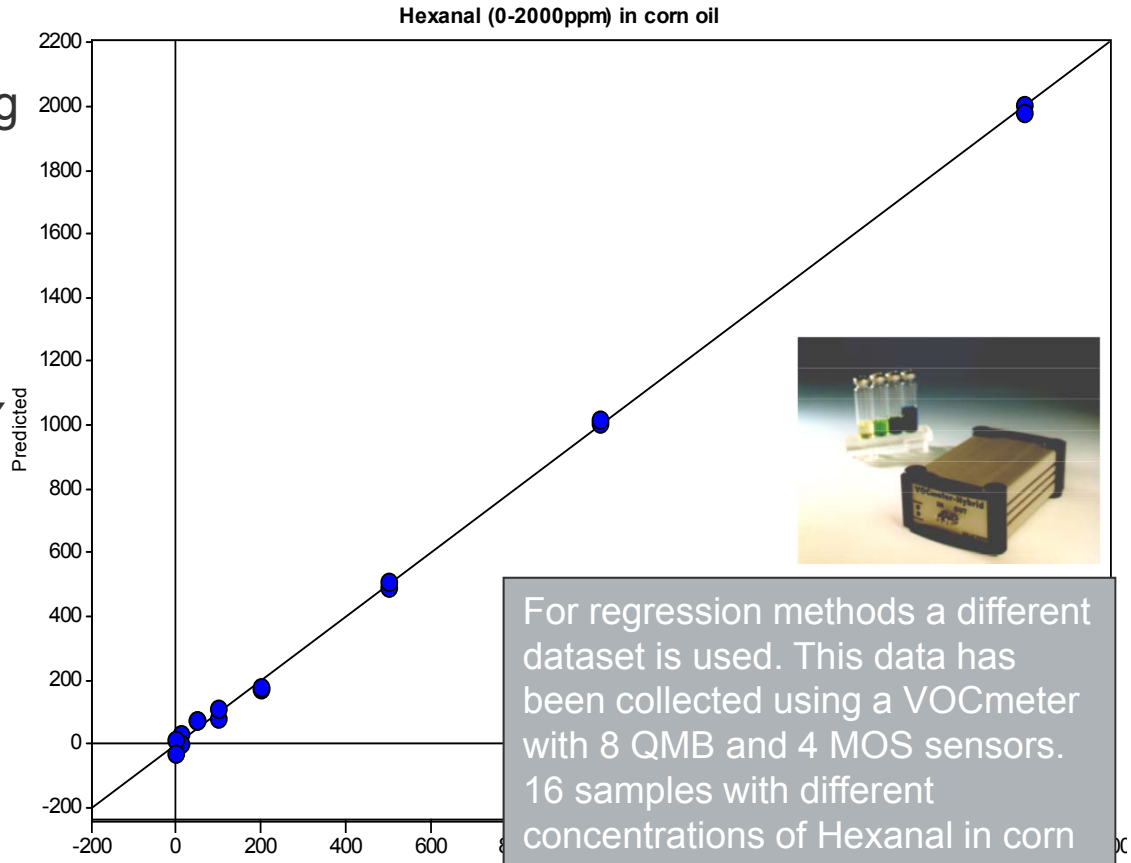
m – number of measurements of y_0

n – number of measurements for regression line

© 2012 University of Tuebingen

Multivariate Linear Regression (MLR): Basic concept

- Prediction of target values Y from X, using a multi-dimensional Least Squares Fit.
- Quantitative relation-ship between X and Y
- The sum of the squared prediction errors of the training set is minimized.



For regression methods a different dataset is used. This data has been collected using a VOCmeter with 8 QMB and 4 MOS sensors. 16 samples with different concentrations of Hexanal in corn oil are analyzed.

Makes use of all available information

 © 2012 University of Tuebingen

MLR: Mathematics

- In MLR we relate **Y** to **X**

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

- Where:

  X is the matrix containing the independent variables
  Y is the matrix containing the dependent variables (true data) for the objects,
  B is a matrix with regression parameters,
  E is a matrix containing the residuals (differences between true and predicted data), that means **Y-XB**

  The least squares fit to calculate the regression parameters is given by:

$$\mathbf{B} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

- Basically, this equation can directly be solved by direct Inversion of the matrix $\mathbf{X^TX}$. In practice $\mathbf{X^TX}$ is often ill conditioned, since the variables are often highly correlated.

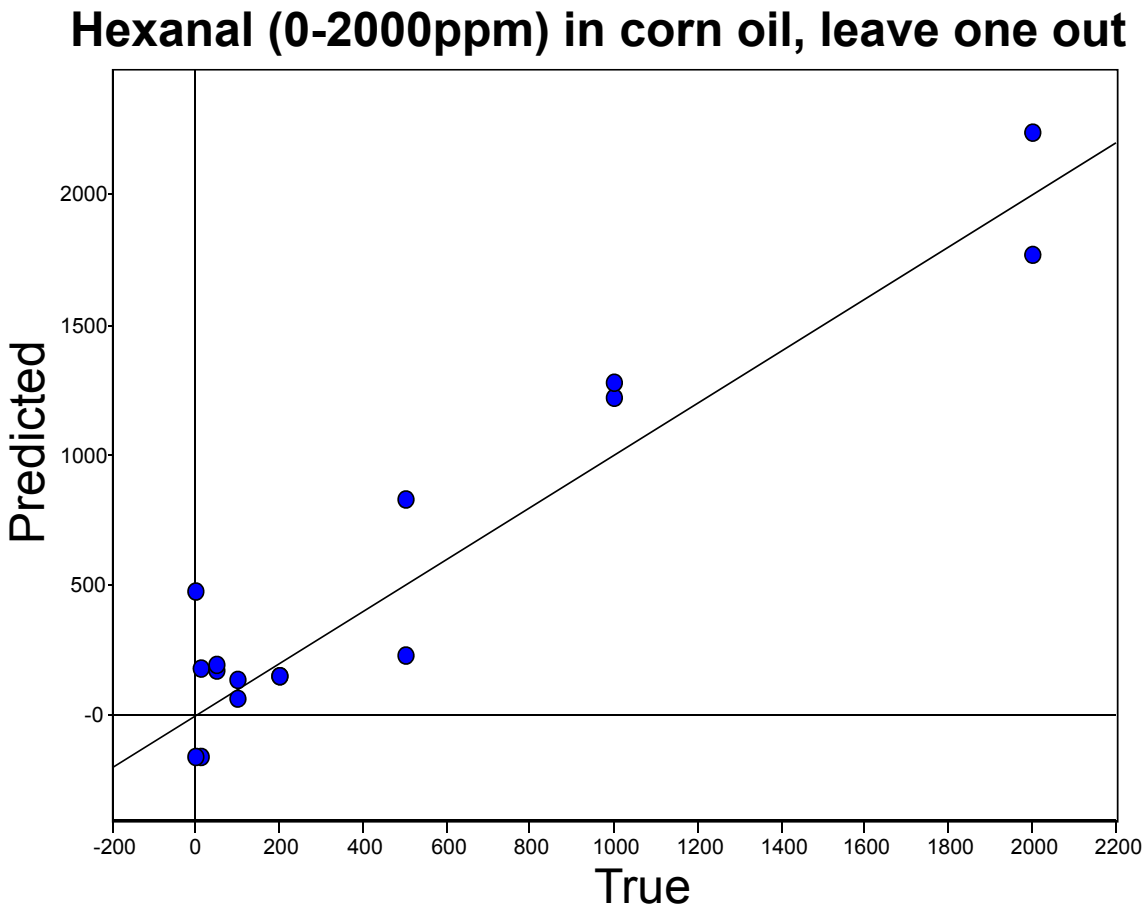
General regression task:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n + e_y$$
$$y = Xb + e_y$$

© 2012 University of Tuebingen

MLR: Overfitting / leave-one-out-cross validation

- In this graph the predicted values are calculated using the leave one out (loo) validation:
- For each object, a model is built using all other objects.
- The prediction error is extremely large!



To reduce the prediction error with MLR, either increase the number of objects or decrease the number of variables / features…

 © 2012 University of Tuebingen

MLR: What is it good for / limitations!

- MLR is good for:

  Regression of data with no correlation within the X-Variables.

- MLR limitations:

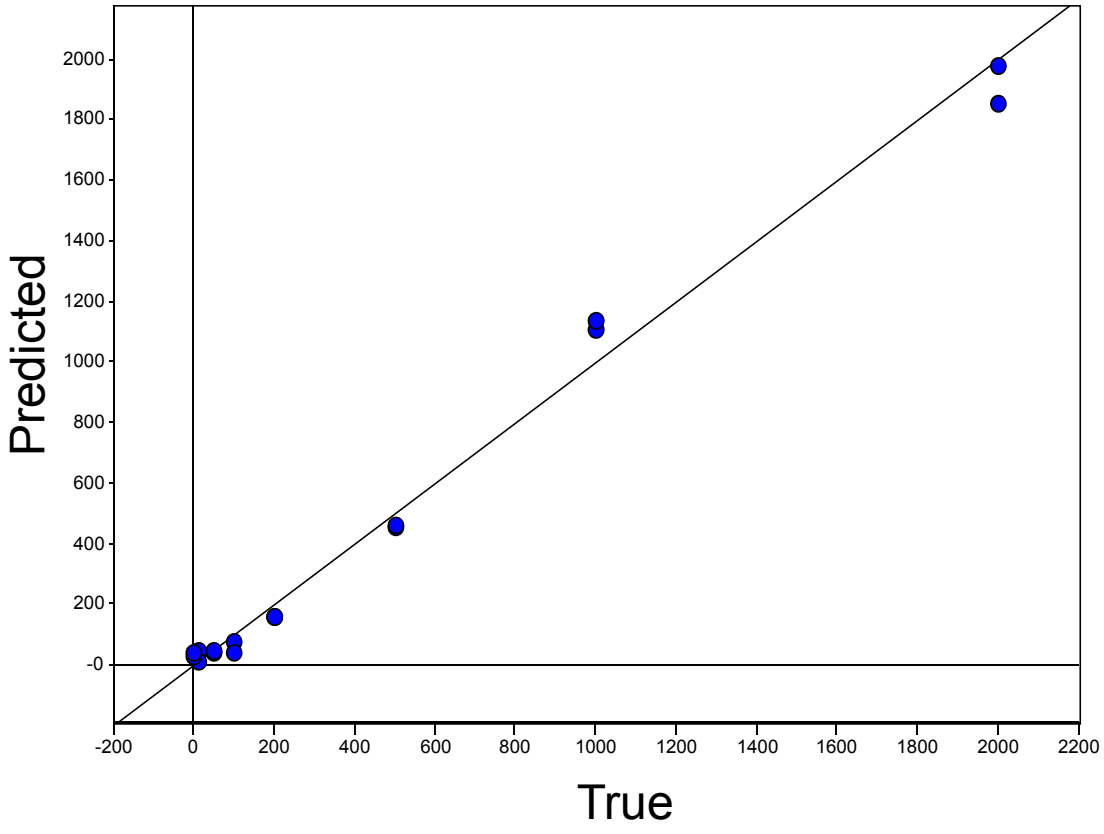
  Correlation is practically always present in analytical data and especially chemical sensor data!

© 2012 University of Tuebingen

Principle Component Regression (PCR): Basic concept

- Extends MLR by introducing a PCA to calculate a new set of variables (principal components), which are then used as input for a MLR.
- The number (and eventually selection) of principal components is an important parameter.
- If all principal components are used, the result is identical to MLR.

Hexanal (0-2000ppm) in corn oil, leave one out, 3 PC's used



Weighting of input variables is an issue since it affects the principal components.

 © 2012 University of Tuebingen

PCR: Calculations

- In PCR we are replacing X by its Singular Value Decomposition (SVD) (see PCA: Calculations):

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}}$$

- We now have the choice to disregard those principal components, that provide only little or noisy information.

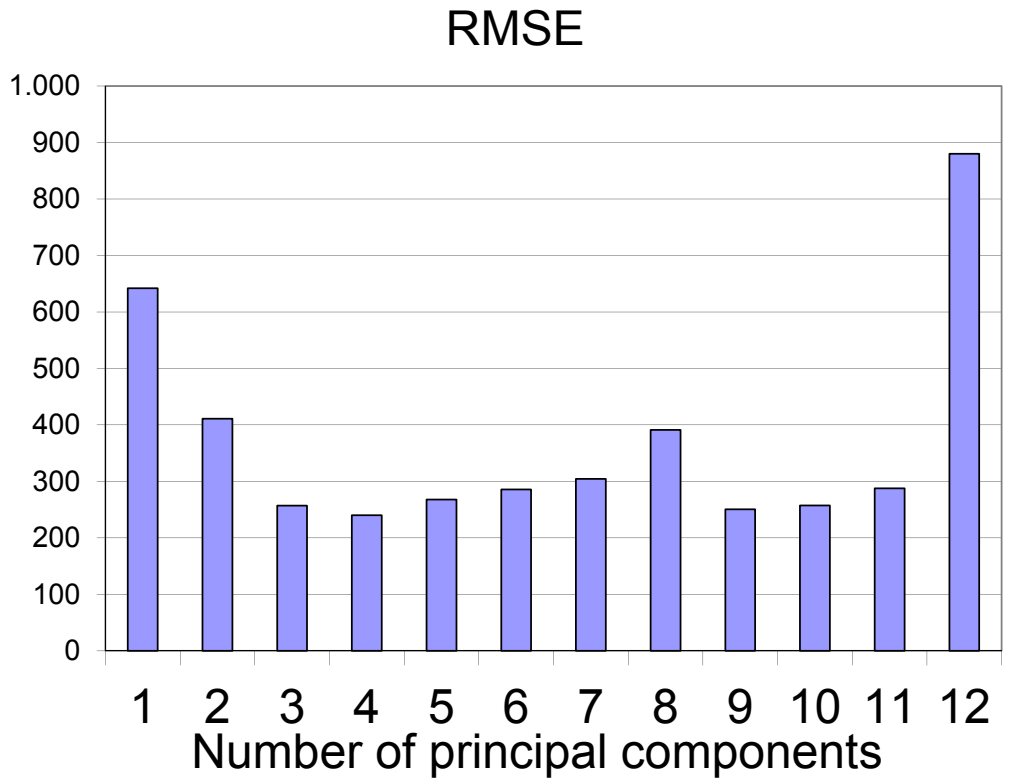
- The parameter matrix **B** for the complete regression is then given by the pseudo-inverse matrix $\mathbf{X}^{+}$ multiplied with the **Y** matrix:

$$\mathbf{B} = \mathbf{X}^{+}\mathbf{Y}, \quad \mathbf{X}^{+} = \mathbf{V}\ \mathbf{diag}\left(1/\sqrt{\lambda_i}\right)\mathbf{U}^{\mathbf{T}}$$

© 2012 University of Tuebingen

PCR: $RMSE_{loo}$ via number of components

- Already 3 principal components are enough to model the data
- When all components are used, the RMSE error using leave one out (loo) increases drastically (equivalent to MLR)



 © 2012 University of Tuebingen

PCR: What is it good for / limitations

- Use PCR for:

  Regression on data, where
  - A limiting set of calibration data is available
  - The desired information can be found (visually) in the scores plot

- Limitations of PCR:
  - The model building does not optimize for the desired information.
  - Nonlinear data may lead to problems.

© 2012 University of Tuebingen

Literature

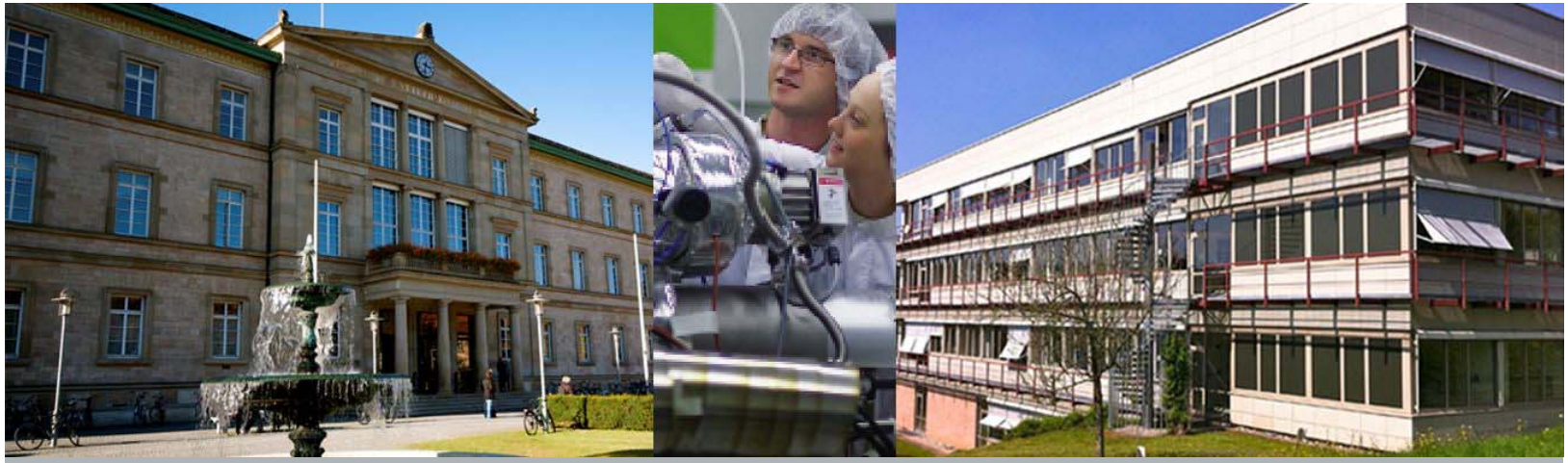
- M. Otto, Chemometrics, Wiley-VCH Weinheim, 1999
- R. Henrion, G. Henrion, Multivariate Datenanalyse, Springer Verlag Berlin, 1994
- H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, 1989
- J. Mitrovics, Auswerteverfahren für Gassensorarrays, Dissertation, 2004
- J. N. Miller, J.C. Miller, Statistics and chemometrics for analytical chemistry, Pearson Education limited, 4th edition, 2000

© 2012 University of Tuebingen

Examination relevant topics

- No mathematical details but an understanding for
  - MLR basic concept
  - PCR basic concept

© 2012 University of Tuebingen



END of Part 04

 © 2012 University of Tuebingen