

## Experiences after 25 years use of multivariate data analysis - chemometrics

NMKL

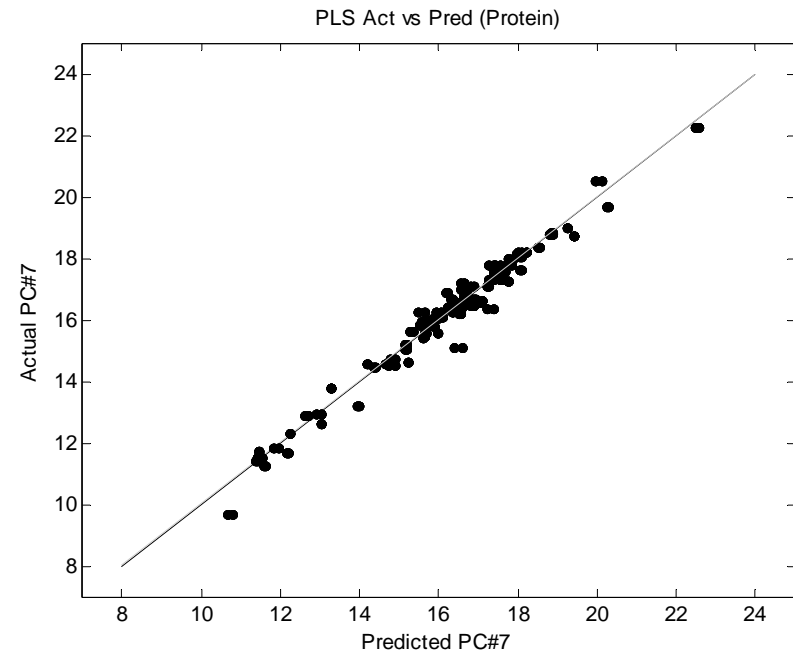
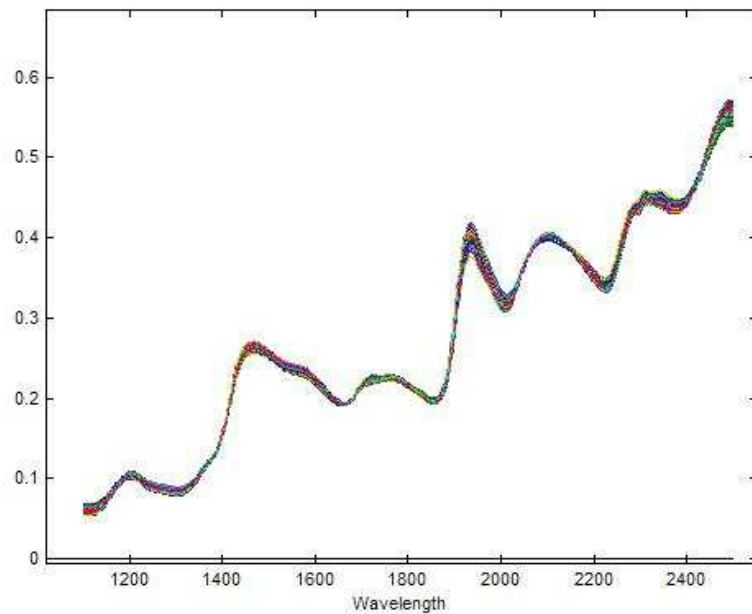
Oslo, August 24 2006

Lars Nørgaard, [lan@kvl.dk](mailto:lan@kvl.dk)

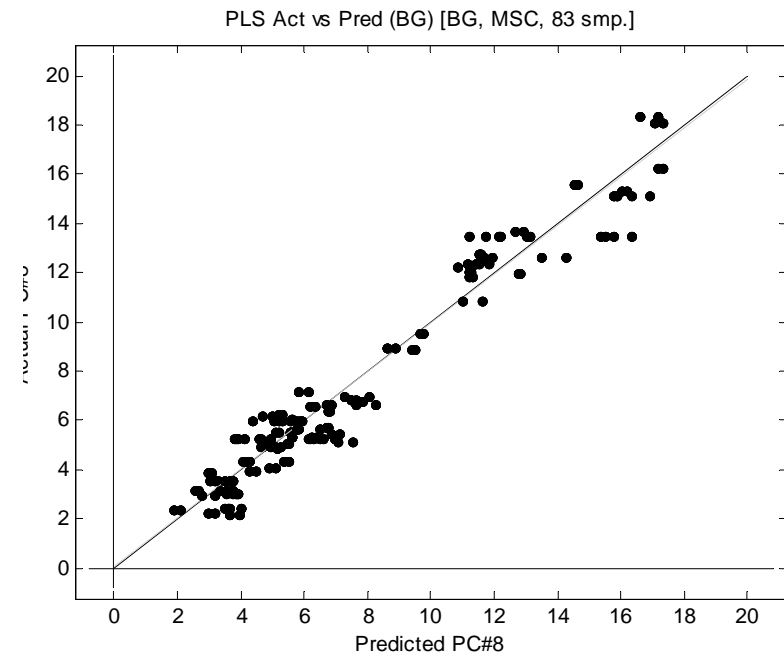
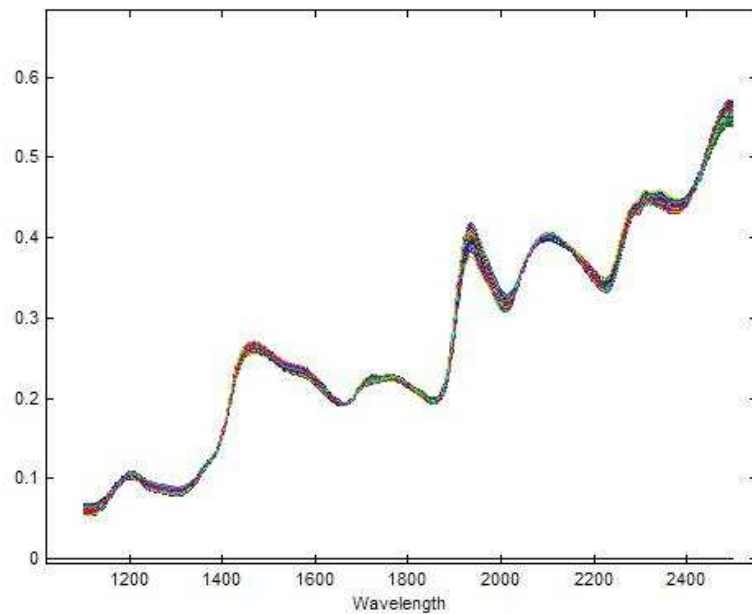
The Royal Veterinary & Agricultural University



# NIR spectra of barley – are these correlated to protein?



## NIR spectra of barley – are these correlated to beta-glucan?



**That's chemometrics:**

**brain + eyes = computer\* + sensors**

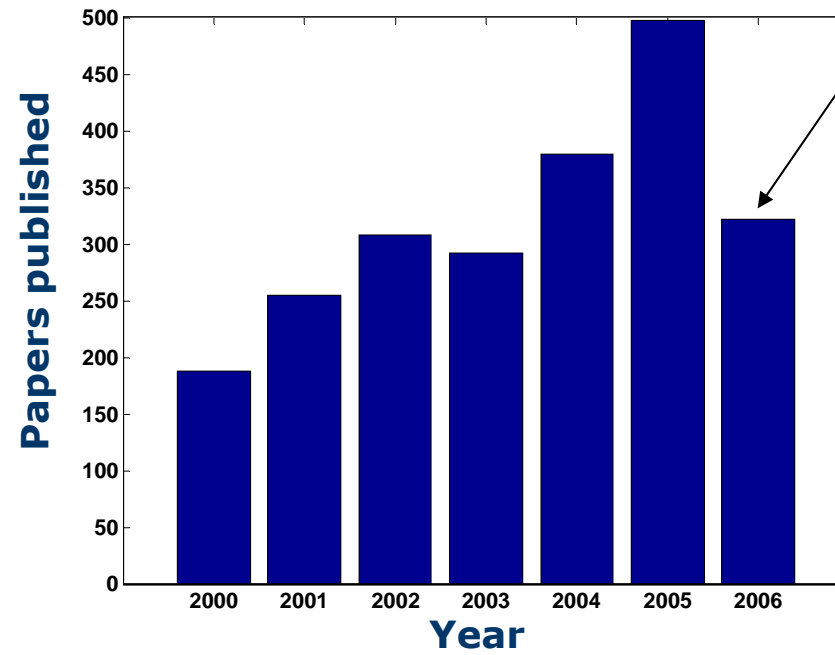
**\*with chemometric software**

## First experience: IT WORKS!

*spectroscopy AND (PCA OR PLS OR "multivariate analysis" OR chemometri\*)*

3145 hits in Web of Science

**By August 21st**



## Also in industry: MilkoScan & WineScan from Foss

The MilkoScan FT120 in the basic configuration offers determination of fat, protein, lactose, total solids and solids-non-fat in milk, cream and simple dairy products.

The WineScan FT120 (Basic) analyses main wine parameters such as ethanol, pH, sugars and organic acids.



## Another experiment...:

PURPOSE:

To show that **Chemometrics = Chemomagics**

Prediction of random numbers from near infrared transmission spectra...

## **Very important experience: ALWAYS VALIDATE YOUR MODELS**

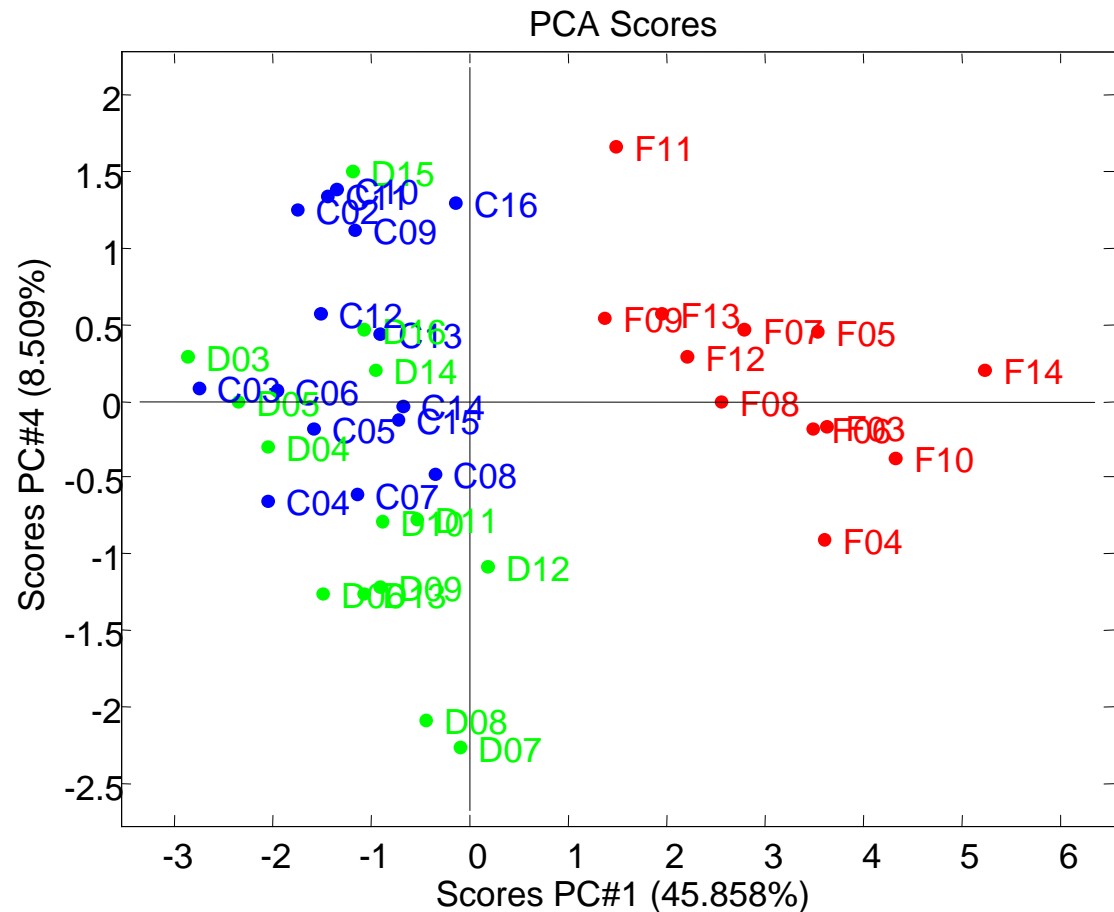
- 1) The optimal: Validation by an external sample set
- 2) If you can't afford it: Validation by an external sample set
- 3) If you still can't afford it: Cross Validation

## How about p-values?

Definitely on the top-ten list of *frequently asked questions* when teaching chemometrics!

Do you need p-values to evaluate this:

If 'Yes' then do it on the scores

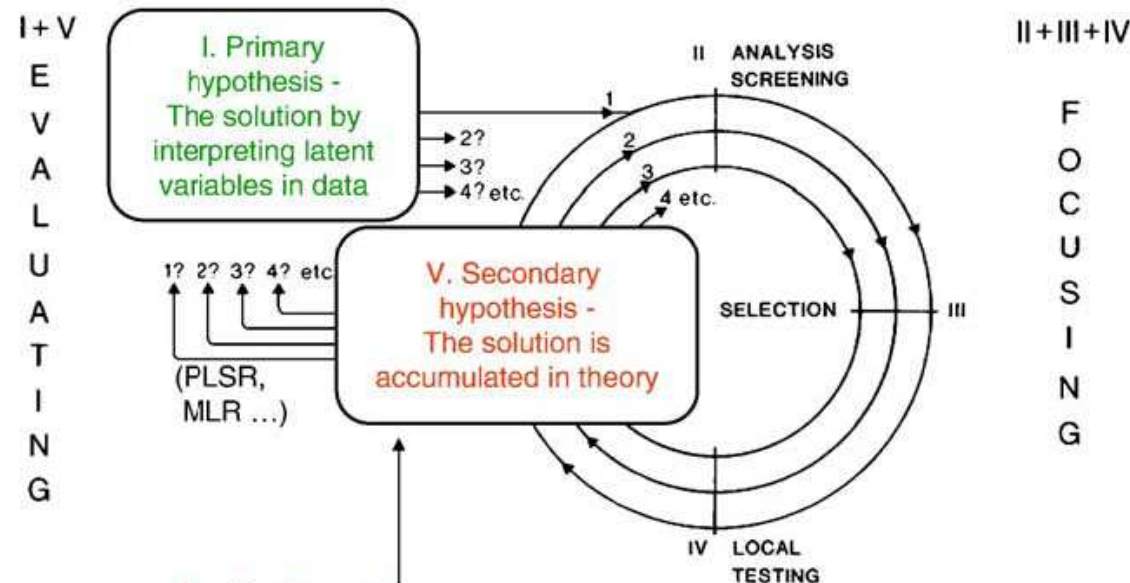


# Chemometrics for exploratory data analysis

## The Latent Variable (LV) Selection Cycle

*Outside the system:*

Surveying by non-destructive screening methods - "top-down modelling" (PCA, PARAFAC ...)



*Inside the system:*

Focusing by destructive analysis - "bottom-up modelling" (Path modelling, PCA, PARAFAC ...)



Munck L, *Conceptual validation of self-organisation studied by spectroscopy in an endosperm gene model as a data-driven logistic strategy in chemometrics*, Chemometrics & Intelligent Laboratory Systems, 2006



# Experience: Visualisation, visualisation and visualisation of methods and data/results

## PCA +25 YEARS AGO:

114 STATISTICAL COMPOSITION 11

The function  $(\mathbf{X} - \lambda)\mathbf{I}$  is a polynomial in  $\lambda$  of degree  $p$ . Therefore (3) has  $p$  roots, let these be  $\lambda_1, \lambda_2, \dots, \lambda_p$ . ( $\mathbf{X}$  complex conjugate in (3) gives  $\lambda$  real.) If we multiply (4) on the left by  $\mathbf{X}$  we obtain

$$(6) \quad \mathbf{X}\mathbf{X}\mathbf{I} = \lambda\mathbf{I}\mathbf{I} = \lambda\mathbf{I}.$$

This shows that  $\mathbf{X}\mathbf{I}$  satisfies (4) (and  $\mathbf{X}\mathbf{I}\mathbf{I} = \lambda\mathbf{I}$ ), then the variance of  $\mathbf{X}\mathbf{I}$  (given by (7)) is  $\lambda$ . Thus for the maximum variance we should use in (4) the largest root  $\lambda_1$ . Let  $\mathbf{I}^1$  be a normalized solution of  $(\mathbf{X} - \lambda_1)\mathbf{I} = \mathbf{0}$ . Then  $\mathbf{I}^1 = \mathbf{X}^1\mathbf{I}$  is a normalized linear combination with maximum variance. If  $(\mathbf{X} - \lambda_1)\mathbf{I}$  is of rank  $p - 1$ , then there is only one solution to  $(\mathbf{X} - \lambda_1)\mathbf{I} = \mathbf{0}$  and  $\mathbf{X}\mathbf{I} = \lambda_1\mathbf{I}$ .

Now let us find a normalized combination  $\mathbf{X}\mathbf{I}$  that has maximum variance of all linear combinations uncorrelated with  $\mathbf{I}^1$ . Lack of correlation is

$$(7) \quad 0 = \mathbf{I}^1\mathbf{X}\mathbf{I} = \mathbf{I}^1\mathbf{X}\mathbf{I}^1\mathbf{I}^1 = \mathbf{I}^1\mathbf{X}\mathbf{I}^1 = \lambda_1\mathbf{I}^1\mathbf{I}^1,$$

since  $\mathbf{X}\mathbf{I}^1 = \lambda_1\mathbf{I}^1$ . Thus  $\mathbf{X}\mathbf{I}$  is orthogonal to  $\mathbf{I}^1$  in both the statistical sense (of lack of correlation) and in the geometric sense (of the inner product of the vectors  $\mathbf{I}$  and  $\mathbf{I}^1$  being zero). (That is,  $\lambda_1\mathbf{I}^1\mathbf{I}^1 = 0$  only if  $\mathbf{I}^1\mathbf{I}^1 = 0$  when  $\lambda_1 \neq 0$ , and  $\lambda_1 = 0$  if  $\mathbf{I}^1 = \mathbf{0}$ ; the case of  $\mathbf{I}^1 = \mathbf{0}$  is obviously trivial and is not treated.) We now want to maximize

$$(8) \quad \lambda_1 = \mathbf{I}^1\mathbf{X}\mathbf{I} = \lambda_1\mathbf{I}^1\mathbf{I} = \lambda_1\mathbf{I}^1\mathbf{X}\mathbf{I}^1,$$

where  $\lambda$  and  $\mathbf{v}_1$  are Lagrange multipliers. The vector of partial derivatives is

$$(9) \quad \frac{\partial \lambda_1}{\partial \mathbf{I}} = \mathbf{X}\mathbf{I} - \lambda\mathbf{I} - \lambda_1\mathbf{X}\mathbf{I}^1,$$

and we set this equal to 0. From (9) we obtain by multiplying on the left by  $\mathbf{I}^1$

$$(10) \quad 0 = \mathbf{I}^1\mathbf{X}\mathbf{I} - \lambda\mathbf{I}^1\mathbf{I} - \lambda_1\mathbf{I}^1\mathbf{X}\mathbf{I}^1 = -2\lambda_1\lambda_1.$$

by (7). Therefore,  $\mathbf{v}_1 = 0$  and  $\mathbf{I}$  must satisfy (4), and therefore  $\lambda$  must satisfy (5). Let  $\lambda_2$  be the maximum of  $\lambda_1, \dots, \lambda_p$ , such that there is a vector  $\mathbf{I}$  satisfying  $(\mathbf{X} - \lambda_2)\mathbf{I} = \mathbf{0}$ ,  $\mathbf{I}\mathbf{I} = 1$ , and (7); call this vector  $\mathbf{I}^2$  and the corresponding linear combination  $\mathbf{I}^2 = \mathbf{X}^2\mathbf{I}$ . (It will be shown eventually that  $\lambda_2 = \lambda_1$ . We define  $\lambda_2 = \lambda_1$ .)

This procedure is continued, at the  $(r + 1)$ st step we want to find a

115 PRINCIPAL COMPONENTS IN THE POPULATION 115

vector  $\mathbf{I}$  such that  $\mathbf{X}\mathbf{I}$  has maximum variance of all normalized linear combinations which are uncorrelated with  $\mathbf{I}^1, \dots, \mathbf{I}^r$ , that is, such that

$$(11) \quad 0 = \mathbf{I}^j\mathbf{X}\mathbf{I} = \mathbf{I}^j\mathbf{X}\mathbf{I}^1\mathbf{I}^1 = \mathbf{I}^j\mathbf{X}\mathbf{I}^1 = \lambda_{j1}\mathbf{I}^j\mathbf{I}^1,$$

We want to maximize

$$(12) \quad \lambda_{j1} = \mathbf{I}^j\mathbf{X}\mathbf{I} = \lambda_{j1}\mathbf{I}^j\mathbf{I} = \lambda_{j1}\mathbf{I}^j\mathbf{X}\mathbf{I}^1,$$

where  $\lambda$  and  $\mathbf{v}_1, \dots, \mathbf{v}_r$  are Lagrange multipliers. The vector of partial derivatives is

$$(13) \quad \frac{\partial \lambda_{j1}}{\partial \mathbf{I}} = \mathbf{X}\mathbf{I} - \lambda\mathbf{I} - 2\sum_{i=1}^r \mathbf{v}_i\mathbf{I}^i\mathbf{I}^i,$$

and we set this equal to 0. Multiplying (13) on the left by  $\mathbf{I}^j$ , we obtain

$$(14) \quad 0 = \mathbf{I}^j\mathbf{X}\mathbf{I} - \lambda\mathbf{I}^j\mathbf{I} - 2\mathbf{v}_j\mathbf{I}^j\mathbf{I}^j.$$

If  $\lambda_{j1} \neq 0$ , this gives  $-2\mathbf{v}_j\lambda_{j1} = 0$  and  $\mathbf{v}_j = 0$ . If  $\lambda_{j1} = 0$ ,  $\mathbf{X}\mathbf{I}^1 = \lambda_{j1}\mathbf{I}^1 = \mathbf{0}$  and the  $j$ th term in the sum in (13) vanishes. Thus  $\mathbf{I}$  must satisfy (4) and therefore  $\lambda$  must satisfy (5).

Let  $\lambda_{m+1}$  be the maximum of  $\lambda_1, \dots, \lambda_p$  such that there is a vector  $\mathbf{I}$  satisfying  $(\mathbf{X} - \lambda_{m+1})\mathbf{I} = \mathbf{0}$ ,  $\mathbf{I}\mathbf{I} = 1$  and (11); call this vector  $\mathbf{I}^{m+1}$ , and the corresponding linear combination  $\mathbf{I}^{m+1} = \mathbf{X}^{m+1}\mathbf{I}$ . If  $\lambda_{m+1} = 0$  and  $\lambda_{m+1} = 0$  ( $j \neq r + 1$ ), then  $\mathbf{I}^{m+1}\mathbf{X}\mathbf{I}^1 = 0$  does not imply  $\mathbf{I}^{m+1}\mathbf{I}^1 = 0$ . However,  $\mathbf{I}^{m+1}$  can be replaced by a linear combination of  $\mathbf{I}^{m+1}$  and the  $\mathbf{I}^j$ 's with  $\lambda_{m+1}$  being 0 so that the new  $\mathbf{I}^{m+1}$  is orthogonal to all  $\mathbf{I}^j$  ( $j = 1, \dots, r$ ). This procedure is carried out until at the  $(m + 1)$ st stage one cannot find a vector  $\mathbf{I}$  satisfying  $\mathbf{I}\mathbf{I} = 1$ , (4), and (11). Either  $m = p$  or  $m < p$  since  $\mathbf{I}^1, \dots, \mathbf{I}^m$  must be linearly independent.

We shall now show that the inequality  $m < p$  leads to a contradiction. If  $m < p$  there exist  $p - m$  vectors, say  $\mathbf{e}_1, \dots, \mathbf{e}_{p-m}$  such that  $\mathbf{I}^j\mathbf{e}_i = 0$ ,  $\mathbf{e}_i\mathbf{e}_i = \lambda_{m+1}$ . (This follows from Lemma 1 in Appendix 1.) Let  $(\mathbf{e}_1, \dots, \mathbf{e}_{p-m}) = \mathbf{E}$ . Now we shall show that there exists a  $(p - m)$ -component vector  $\mathbf{c}$  and a number  $\beta$  such that  $\mathbf{E}\mathbf{c} = \beta\mathbf{E}$  is a solution to (4) with  $\lambda = 0$ . Consider a root of  $(\mathbf{E}\mathbf{E} - \lambda)\mathbf{I} = \mathbf{0}$  and a corresponding vector  $\mathbf{c}$  satisfying  $\mathbf{E}\mathbf{E}\mathbf{c} = \lambda\mathbf{c}$ . The vector  $\mathbf{E}\mathbf{c}$  is orthogonal to  $\mathbf{I}^1, \dots, \mathbf{I}^m$  (since  $\mathbf{I}^j\mathbf{E}\mathbf{c} = \lambda_{j1}\mathbf{I}^j\mathbf{c} = \lambda_{j1}\mathbf{I}^j\mathbf{I}^j\mathbf{c} = 0$ ) and therefore is a vector in the space spanned by  $\mathbf{e}_1, \dots, \mathbf{e}_{p-m}$  and can be written as  $\mathbf{E}\mathbf{g}$  (where  $\mathbf{g}$  is a  $(p - m)$ -component vector). Multiplying  $\mathbf{E}\mathbf{E}\mathbf{c} = \mathbf{E}\mathbf{g}$  on the left by  $\mathbf{E}$ , we obtain  $\mathbf{E}\mathbf{E}\mathbf{E}\mathbf{c} = \mathbf{E}\mathbf{E}\mathbf{g}$ . Thus  $\mathbf{g} = \mathbf{0}$  and we have  $\mathbf{E}\mathbf{c} = \mathbf{0}$ . Then  $(\mathbf{E}\mathbf{c})\mathbf{X}$  is uncorrelated with  $\mathbf{I}^1, \dots, \mathbf{I}^m$  and this leads to a new  $\mathbf{I}^{m+1}$ . Since this contradicts the assumption that  $m < p$ , we must have  $m = p$ .



## PCA today

Director:  
Rasmus Bro

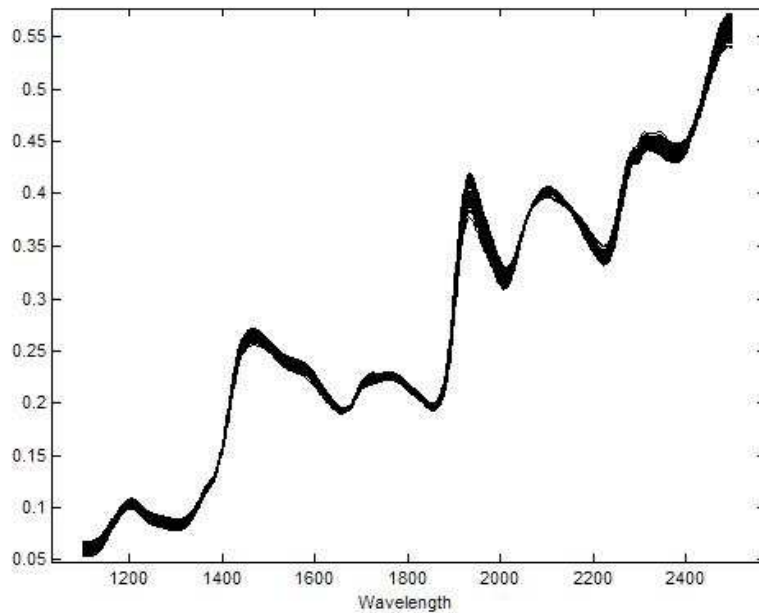
[www.models.kvl.dk](http://www.models.kvl.dk)



	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

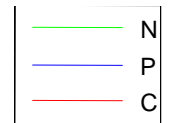
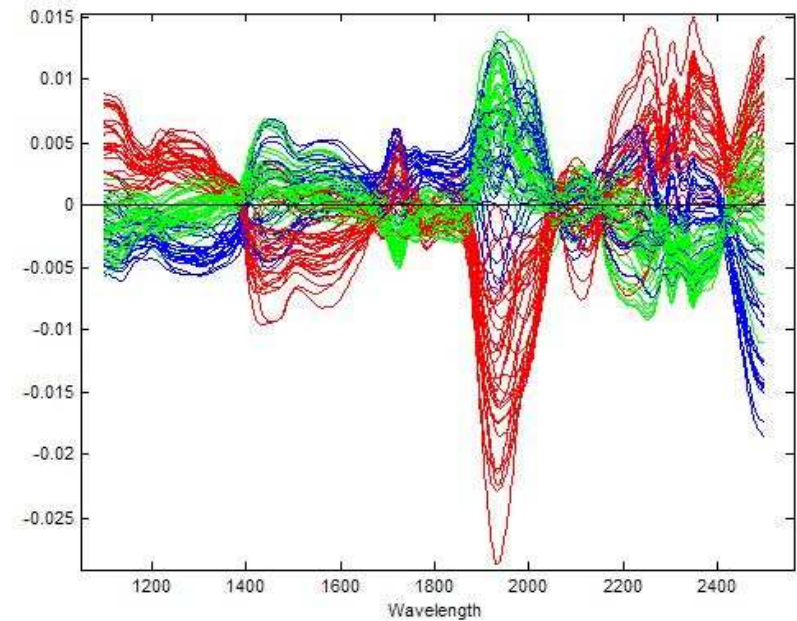
# Raw data 25 years ago

Raw NIR spectra of barley samples



# and today

MSC+mean cent. NIR spectra of barley samples



## Experience: in order to make method development co-operation is a must!

FDA has recognized that!

### Guidance for Industry PAT – A Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance

*Additional copies are available from:*

*Office of Training and Communication  
Division of Drug Information, HFD-240  
Center for Drug Evaluation and Research  
Food and Drug Administration  
5600 Fishers Lane  
Rockville, MD 20857  
(Tel) 301-827-4573  
<http://www.fda.gov/cder/guidance/index.htm>*

*and/or*

*Communications Staff, HFV-12  
Center for Veterinary Medicine  
Food and Drug Administration  
7519 Standish Place,  
Rockville, MD 20855  
(Tel) 301-827-3800  
<http://www.fda.gov/cvm/guidance/published.html>*

**U.S. Department of Health and Human Services  
Food and Drug Administration  
Center for Drug Evaluation and Research (CDER)  
Center for Veterinary Medicine (CVM)  
Office of Regulatory Affairs (ORA)  
August 2003  
Pharmaceutical CGMPs**

## Old and new ways

### Current operating scenario

Products are manufactured according to fixed process conditions set during development and confirmed during initial process and product validation. Release is conducted by physical and chemical testing subsequent to manufacture.

*”From a physical, chemical or biological perspective **food and feed** products and processes are complex multi-factorial systems”*

### New approach

Product is manufactured according to process conditions that are responding to direct measurement of in-process product quality. Relationships are developed between process and product performance. Release is by data collected from in-process product or each dosage form during manufacture.

## Why?

"The term analytical in PAT is viewed broadly to include chemical, physical, microbiological, mathematical, and risk analysis conducted in an integrated manner. The goal of PAT is to understand and control the manufacturing process:

**Quality cannot be tested into products; it should be built-in or should be by design."**

=====

"Data analysis and management is used for obtaining quality, not to report it"



Bookmark	Scope
Pages	Referenced Docu
	Terminology
	Summary of Prac
	Significance and
	Overview of Multiv
	Infrared Instrumen
	Infrared Spectral
	Reference Method
	Simple Procedure
	Data Preprocessi
	Multivariate Calibr
	Estimation of Valu
	Post Processing
	Statistics Used in
	Outlier Statistics
	Selection of Calib
	Validation of a Mu
	Precision of Infra
	Major Sources of
	Wavelength (Freq
	TABLE 1
	Calibration Transf
	TABLE 2
	TABLE 3
	TABLE 4
	Calibration Quality
	Model Updating
	Multivariate Calibr
	Keywords
Comments	A1. STATISTICAL
	A1.1 Dixon's Test
	A1.2



Designation: E 1655 – 05

## Standard Practices for Infrared Multivariate Quantitative Analysis<sup>1</sup>

This standard is issued under the fixed designation E 1655; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

### 1. Scope

1.1 These practices cover a guide for the multivariate calibration of infrared spectrometers used in determining the physical or chemical characteristics of materials. These practices are applicable to analyses conducted in the near infrared (NIR) spectral region (roughly 780 to 2500 nm) through the mid infrared (MIR) spectral region (roughly 4000 to 400  $\text{cm}^{-1}$ ).

NOTE 1—While the practices described herein deal specifically with mid- and near-infrared analysis, much of the mathematical and procedural detail contained herein is also applicable for multivariate quantitative analysis done using other forms of spectroscopy. The user is cautioned that typical and best practices for multivariate quantitative analysis using other forms of spectroscopy may differ from practices described herein for mid- and near-infrared spectroscopies.

1.2 Procedures for collecting and treating data for developing IR calibrations are outlined. Definitions, terms, and calibration techniques are described. Criteria for validating the performance of the calibration model are described.

1.3 The implementation of these practices require that the IR spectrometer has been installed in compliance with the manufacturer's specifications. In addition, it assumes that, at the times of calibration and of validation, the analyzer is operating at the conditions specified by the manufacturer.

of the multivariate mathematics described herein, they do not conform to procedures described herein, specifically with respect to the handling of outliers. Surrogate methods may indicate that they make use of the mathematics described herein, but they should not claim to follow the procedures described herein.

1.7 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.*

### 2. Referenced Documents

#### 2.1 ASTM Standards:<sup>2</sup>

- D 1265 Practice for Sampling Liquefied Petroleum (LP) Gases (Manual Method)
- D 4057 Practice for Manual Sampling of Petroleum and Petroleum Products
- D 4177 Practice for Automatic Sampling of Petroleum and Petroleum Products
- D 4855 Practices for Comparing Test Methods
- D 6122 Practice for Validation of Multivariate Process Infrared Spectrophotometers
- D 6299 Practice for Applying Statistical Quality Assurance

## Checking the two journals on chemometrics

### SEARCHING WEB OF SCIENCE:

*Journal of Chemometrics OR Chemometrics and Intelligent Laboratory Systems*

2644 hits for these journals

## Checking the two journals on chemometrics

1. Wold S, Esbensen K, Geladi P, [Principal Component Analysis](#), **CHEMOLAB**, 1987  
**Times Cited:** [959](#)
2. Bro R, [PARAFAC. Tutorial and applications](#), **CHEMOLAB**, 1997  
**Times Cited:** [276](#)
3. Tauler R, Smilde A, Kowalski B, [Selectivity, Local Rank, 3-Way Data ...](#), **J CHEMOMETRICS**, 1995  
**Times Cited:** [221](#)
4. Wold S, Antti H, Lindgren F, et al. [Orthogonal signal correction of NIR spectra](#), **CHEMOLAB**, 1998  
**Times Cited:** [199](#)
5. Wold S, Kettaneh-Wold N, Skagerberg B, [Non-linear PLS Modeling](#), **CHEMOLAB**, 1989  
**Times Cited:** [198](#)
6. Wold S, Sjostrom M, Eriksson L, [PLS-regression: a basic tool of chemometrics](#), **CHEMOLAB**, 2001  
**Times Cited:** [172](#)
7. Leardi R, Boggia R, Terrile M, [Genetic Algorithms as a Strategy...](#), **J CHEMOMETRICS**, 1992  
**Times Cited:** [170](#)
8. De Jong S, [SIMPLS - An Alternative Approach to PLS Regression](#), **CHEMOLAB**, 1993  
**Times Cited:** [166](#)
9. Bro R, [Multiway calibration. Multilinear PLS](#), **J CHEMOMETRICS**, 1996  
**Times Cited:** [160](#)
10. Lucasius CB, Kateman G, [Understanding and Using Genetic Algorithms...](#), **CHEMOLAB**, 1993  
**Times Cited:** [149](#)



## Checking other journals and a book...

MARTENS H, NAES T [MULTIVARIATE CALIBRATION](#) WILEY, 1989  
Times Cited: MORE THAN [2000](#)

GELADI P, KOWALSKI BR [PARTIAL LEAST-SQUARES REGRESSION - A TUTORIAL](#) ANAL. CHIM. ACTA  
Times Cited: [1327](#)

HAALAND DM, THOMAS EV, [PARTIAL LEAST-SQUARES METHODS FOR SPECTRAL ANALYSIS](#) 1988  
Times Cited: [917](#)

WOLD S, RUHE A, WOLD H, et al. [THE COLLINEARITY PROBLEM IN LINEAR-PARTIAL LEAST-SQUARES \(PLS\) APPROACH TO GENERALIZED INVERSES](#), SIAM J SCIENTIFIC & STAT. COMPUTING  
Times Cited: [430](#)

GELADI P, MACDOUGALL D, MARTENS H [LINEARIZATION AND SMOOTHING OF NEAR-INFRARED REFLECTANCE SPECTRA OF MEAT](#), APPLIED SPECTROSCOPY 1985  
Times Cited: [315](#)

FRANK IE, FRIEDMAN JH, [A STATISTICAL VIEW OF REGRESSION ANALYSIS](#) REGRESSION TOOLS, TECHNOMETRICS 1993  
Times Cited: [283](#)

LINDBERG W, PERSSON JA, WOLD S, [PARTIAL LEAST-SQUARES REGRESSION METHOD FOR SPECTROFLUORIMETRIC ANALYSIS OF MIXTURES OF HUMIC-ACID AND LIGNINSULFONATES](#) JOURNAL OF CHEMICAL METROLOGY 1983  
Times Cited: [276](#)

STONE M, BROOKS RJ [ADAPTIVELY SEQUENTIALLY CONSTRUCTED PREDICTION EMBRACING ORDINARY LEAST-SQUARES, PARTIAL LEAST-SQUARES, AND PRINCIPAL COMPONENTS REGRESSION](#), JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-METHODS AND STATISTICS 1992  
Times Cited: [1000](#)

ISAKSSON T, NAES T [EFFECT OF MULTIPLICATIVE SCATTER CORRECTION \(MSC\) AND LINEARITY IMPROVEMENT IN NIR SPECTROSCOPY](#) APPLIED SPECTROSCOPY 1996  
Times Cited: [146](#)

**PCA, PLS, MULTI-WAY, PRE-PROCESSING AND VARIABLE SELECTION**

**NOTE: NOT A COMPLETE LIST OF PAPERS**

## A personal view on the multivariate future:

- MODELS for MEGAVARIATE DATA
  - DATA FUSION  
(that's already happening but we are only at the beginning)
- VISUALIZATION of such data (e.g. 100,000 samples)
- DEVELOPING NEW SENSORS/INSTRUMENTS  
Based on the fact that we can deal with multivariate data

## Main conclusion

**Chemometrics is here to stay and it WILL still work in the future...**