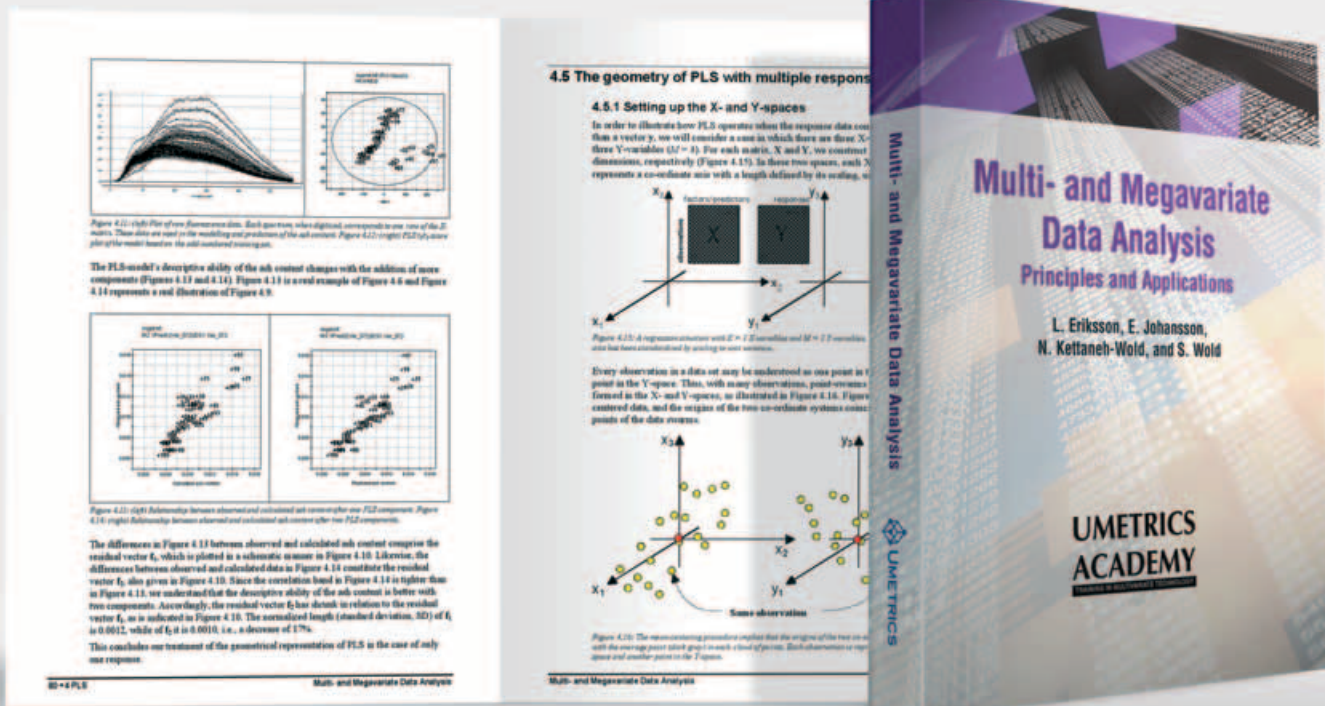


# Multi- and Megavariate Data Analysis

## Principles and Applications

ISBN 91-973730-1-X

L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold



## How to Manage the Data Explosion

“Our new process equipment can read and measure practically everything. But how are we supposed to keep track of all the incoming signals and data?”

Lack of data is hardly the problem these days. With multivariate data analysis, it's possible to manage even the largest datasets and still interpret them quickly and confidently. This 533-page book handles the basic concepts and principles of projections and introduces the two modeling methods, Principal Component Analysis (PCA) and Partial Least Squares (PLS). Different areas of applications are discussed and exemplified with real-life data sets. The authors offer their detailed analyses and offer solutions, with the graphical presentation that is the trademark of Umetrics SIMCA software family.

L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold are leading experts in Multivariate Data Analysis and have a vast experience of application areas from years of consulting and lecturing at Umetrics.

“... in this book the authors do a wonderful job of introducing key concepts graphically with the use of a number of well-chosen examples.”

“In conclusion, the authors have done an excellent job at meeting their stated objective of describing multivariate data analysis tools, namely PCA and PLS, in a practical manner.”

Ronald E. Shaffer in *Journal of Chemometrics*, 16, 2002, 261-262



# Content of “Multi- and Megavariate Data Analysis”

## 0 Preface

- 1 Why MultiVariate Data Analysis (MVDA) ?
- 2 Projection Methods (Principal Component Analysis, PCA, and Projections to Latent Structures, PLS)
- 3 Recognition of differences between multi- and megavariate analysis
- 4 Problems areas where MVDA is useful
- 5 Objectives of the book
- 6 Organization of the book

## 1 Introduction

- 1 Objective
- 2 Introduction
- 3 Pattern recognition
- 4 Basic types of data analytical questions
- 5 Illustrations of problem types – Overview of a data set
- 6 Illustration of problem types – Classification & Discrimination
- 7 Illustration of problem types – Regression modeling

## 2 Basic concepts and principles of projections

- 1 Objective
- 2 Data
- 3 Some properties and challenges of multivariate data
- 4 Confusion of correlation and causation
- 5 How do we analyze our data today?
- 6 The principles of projections
- 7 The model concept and theoretical basis for model building

## 3 PCA

- 1 Objective
- 2 Introduction to PCA
- 3 Pre-treatment of data
- 4 A geometric interpretation of PCA
- 5 Additional PCA diagnostics

## 4 PLS

- 1 Objective
- 2 Introduction to PLS
- 3 Pre-processing of data
- 4 The geometry of PLS in the case of one response ( $M = 1$ )
- 5 The geometry of PLS with multiple responses ( $M > 1$ )
- 6 PLS-model interpretation
- 7 Additional PLS diagnostics
- 8 Use of PLS-model: Predictions

## 5 Multivariate characterization

- 1 Objective
- 2 Introduction
- 3 Rationale of multivariate characterization

- 4 Main steps of multivariate characterization
- 5 Main example: SURFACTANTS

## 6 Multivariate calibration

- 1 Objective
- 2 Introduction
- 3 Multivariate calibration
- 4 Example: SUGAR
- 5 A second example: SOIL\_DNA
- 6 A third example: SAWDUST

## 7 Multivariate process modeling

- 1 Objective
- 2 Introduction
- 3 Multivariate process modeling
- 4 Example: CUPRUM
- 5 A second example: PROC1A
- 6 A third example: SOVRING

## 8 Classification and discrimination

- 1 Objective
- 2 Introduction
- 3 Pattern recognition (PARC)
- 4 Example: IRIS
- 5 Partial least squares discriminant analysis (PLS-DA)
- 6 Example: ARCHAIE

## 9 Transformation and expansion

- 1 Objective
- 2 Transformation
- 3 Example: NEURO
- 4 Expansion of the X-matrix
- 5 Example: SOVRING

## 10 Scaling

- 1 Objective
- 2 Introduction
- 3 Example: LOWARP
- 4 Example: SUGAR
- 5 Related scaling procedures
- 6 Trimming and Winsorizing

## 11 Signal correction and compression

- 1 Objective
- 2 Introduction
- 3 Methods for signal correction and compression
- 4 Example: 34PCBs
- 5 Example: SUGAR
- 6 Example: SUPERSUG

## 12 MSPC

- 1 Objective
- 2 Introduction
- 3 Statistical process control (SPC)
- 4 The Shewhart chart
- 5 The CuSum chart
- 6 The EWMA chart
- 7 Example: PROC1A

## 13 BSPC

- 1 Objective

- 2 Introduction
- 3 A novel approach to BSPC
- 4 Observation level: Modeling batch evolution (EXAMPLE: BYEAST)
- 5 Batch level: Modeling final batch results (EXAMPLE: BYEAST)
- 6 A second batch example: NOM18A

## 14 Multivariate time series analysis

- 1 Objective
- 2 Outline of Chapter 14
- 3 Introduction to time series data
- 4 Review of time series and their properties
- 5 The transfer function and PLS time series analysis
- 6 MTSA and lagging of latent variables (Example: CUPRUM)
- 7 MTSA using PLS (Example: KAMYR)

## 15 DOE in industrial practice

- 1 Objective
- 2 Introduction
- 3 Design of Experiments (DOE)
- 4 A screening case study: ENVIRO
- 5 An optimization case study: SOVRING
- 6 A robustness testing case study: HPLC\_ROB

## 16 A multivariate approach to QSAR

- 1 Objective
- 2 Introduction
- 3 Fundamental conditions of QSAR modeling
- 4 Example 1 – The SCHÜRMANN data set ( $N=20$ ,  $K=5$ ,  $M=1$ )
- 5 Example 2 – The McCLOSKEY data set ( $N=20$ ,  $K=13$ ,  $M=1$ )
- 6 Example 3 – The MÜLLER data set ( $N=66$ ,  $K=56$ ,  $M=1$ )
- 7 Example 4 – The HERMENS data set ( $N=15$ ,  $K=8$ ,  $M=8$ )
- 8 Example 5 – The KARLÉN data set ( $N=22$ ,  $K=18$ ,  $M=1$ )
- 9 Example 6 – The CELLTEST data set ( $N=16$ ,  $K=6$ ,  $M=2$ )

## 17 Peptide QSAR

- 1 Objective
- 2 Introduction
- 3 Example: Z-SCALES
- 4 Example: ANGIO
- 5 Example: BITTER
- 6 Example: PENTAPEP

## 18 Lead finding and optimization

- 1 Objective
- 2 Introduction
- 3 Lead finding using pharmacological profiling (Example: ALDRICH)

- 4 Lead optimization with multivariate design (EXAMPLE: ALDRICH)
- 5 Multivariate modeling and design of hexapeptides (Example: HEXAPEP)
- 6 Multivariate design in constrained PP-spaces (Example: PCBCYP2B)

## 19 Multivariate combinatorial chemistry

- 1 Objective
- 2 Introduction
- 3 Chemical characterization of compounds and building blocks (Step 1)
- 4 Selection of representative building blocks (Step 2a)
- 5 Generation of the final library (Step 2b)
- 6 Biological testing (Step 3)
- 7 Linking of chemical and biological data: QSAR (Step 4)

## 20 Chem- and Bioinformatics

- 1 Objective
- 2 Introduction
- 3 Example: PromSeq
- 4 Example: ACCPEP

## 21 Non-linear PLS-modeling

- 1 Objective
- 2 Introduction
- 3 Binning and expansion of X-variables (GIFI-PLS)
- 4 QSAR Example: ELASTASE (Peptide QSAR)
- 5 Process Example: SimCODM

## 22 Multivariate analysis of preference data

- 1 Objective
- 2 Introduction
- 3 Multivariate conjoint analysis (Example: CONJOINT)
- 4 Multivariate preference mapping (Example: SENSCONS)

## 23 Model derivation, interpretation, and validation

- 1 Objective
- 2 Introduction
- 3 Evaluation and pre-processing of raw data
- 4 Model derivation and interpretation
- 5 Model validation and use of model

## Statistical Appendix

## References

## Index



www.umetrics.com

In all countries:  
**Umetrics AB**  
Box 7960  
SE-90719 Umeå  
Sweden  
Phone: +46 (0)90 184800  
Fax: +46 (0)90 184899  
Email: info.se@umetrics.com

In USA and Canada:  
**Umetrics Inc.**  
17 Kiel Ave.  
Kinnelon NJ 07405  
USA  
Phone: +1 973 492 8355  
Fax: +1 973 492 8359  
Email: info.us@umetrics.com

In UK:  
**Umetrics UK Ltd.**  
Woodside House  
Winkfield, Windsor  
Berkshire, SL4 2DX, UK  
Phone: +44 (0)1344 885605  
Fax: +44 (0)1344 885410  
Email: info.uk@umetrics.com