

Introduction to
**Multivariate
Statistical Analysis
in Chemometrics**

Kurt Varmuza
Peter Filzmoser



 **CRC Press**
Taylor & Francis Group



ISBN: 9781420059472

CRC Press (Taylor & Francis),
Boca Raton, FL, USA, www.crcpress.com

Publication Date: February 2009

Ca 320 pages

Price: ca US\$ 120

Authors

Kurt Varmuza and
Peter Filzmoser

Vienna University of Technology,
Austria

kvarmuza@email.tuwien.ac.at
P.Filzmoser@tuwien.ac.at

- Includes important multivariate statistical methods, such as principal component analysis and support vector machines, for analyzing scientific data
- Explains the methods using formulae, graphical illustrations, and schemes
- Demonstrates **R software tools** with fully worked-out, real-world examples
- Emphasizes the use of **robust statistical methods**
- Offers practical advice on applying the methods

Using formal descriptions, graphical illustrations, practical examples, and R software tools,

Introduction to Multivariate Statistical Analysis in Chemometrics

presents simple yet thorough explanations of the most important multivariate statistical methods for analyzing chemical data.

It includes discussions of various statistical methods, such as principal component analysis, regression analysis, classification methods, and clustering.

Written by a chemometrician and a statistician, the book reflects both the practical approach of chemometrics and the more formally oriented one of statistics. To enable a better understanding of the statistical methods, the authors apply them to real data examples from chemistry. They also examine results of the different methods, comparing traditional approaches with their robust counterparts. In addition, the authors use the freely available R package to implement methods, encouraging readers to go through the examples and adapt the procedures to their own problems.

Focusing on the practicality of the methods and the validity of the results, this book offers concise mathematical descriptions of many multivariate methods and employs graphical schemes to visualize key concepts. It effectively imparts a basic understanding of how to apply statistical methods to multivariate scientific data.

Contents

- 1 Introduction**
 - 1.1 Chemoinformatics - chemometrics - statistics**
 - 1.2 This book**
 - 1.3 Historical remarks about chemometrics**
 - 1.4 Bibliography**
 - 1.5 Starting examples**
 - 1.5.1 Univariate versus bivariate classification
 - 1.5.2 Nitrogen content of cereals computed from NIR data
 - 1.5.3 Elemental composition of aecheaeological glasses
 - 1.6 Univariate statistics - a reminder**
 - 1.6.1 Empirical distributions
 - 1.6.2 Theoretical distributions
 - 1.6.3 Central value
 - 1.6.4 Spread
 - 1.6.5 Statistical tests
- 2 Multivariate data**
 - 2.1 Definitions**
 - 2.2 Basic preprocessing**
 - 2.2.1 Data transformation
 - 2.2.2 Centering and scaling
 - 2.2.3 Normalization
 - 2.2.4 Transformations for compositional data
 - 2.3 Covariance and correlation**
 - 2.3.1 Overview
 - 2.3.2 Estimating covariance and correlation
 - 2.4 Distances and similarities**
 - 2.5 Multivariate outlier identification**
 - 2.6 Linear latent variables**
 - 2.6.1 Overview
 - 2.6.2 Projection and mapping
 - 2.6.3 Example
 - 2.7 Summary**
- 3 Principal component analysis**
 - 3.1 Concepts**
 - 3.2 Number of PCA components**
 - 3.3 Centering and scaling**
 - 3.4 Outliers and data distribution**
 - 3.5 Robust PCA**
 - 3.6 Algorithms for PCA**
 - 3.6.1 Mathematics of PCA
 - 3.6.2 Jacobi rotation
 - 3.6.3 Singular value decomposition
 - 3.6.4 NIPALS
 - 3.7 Evaluation and diagnostics**
 - 3.7.1 Cross validation for determination of the number of principal components
 - 3.7.2 Explained variance for each variable
 - 3.7.3 Diagnostic plots
 - 3.8 Complementary methods for exploratory data analysis**
 - 3.8.1 Factor analysis
 - 3.8.2 Cluster analysis and dendrogram
 - 3.8.3 Kohonen mapping

	3.8.4	Sammon's nonlinear mapping		4.5.5	Variable selection based on PCA or PLS models
	3.8.5	Multi-way PCA		4.5.6	Genetic algorithms
3.9	Examples			4.5.7	Cluster analysis of variables
	3.9.1	Tissue samples from human mummies and fatty acid concentrations		4.5.8	Example
	3.9.2	Polycyclic aromatic hydrocarbons in aerosol	4.6	Principal component regression	
3.10	Summary			4.6.1	Overview
4	Calibration			4.6.2	Number of PCA components
	4.1	Concepts	4.7	Partial least-squares regression	
	4.2	Performance of regression models		4.7.1	Overview
	4.2.1	Overview		4.7.2	Mathematical aspects
	4.2.2	Overfitting and underfitting		4.7.3	Kernel algorithm for PLS
	4.2.3	Performance criteria		4.7.4	NIPALS algorithm for PLS
	4.2.4	Criteria for models with different numbers of variables		4.7.5	SIMPLS algorithm for PLS
	4.2.5	Cross validation		4.7.6	Other algorithms for PLS
	4.2.6	Bootstrap		4.7.7	Robust PLS
	4.3	Ordinary least-squares regression	4.8	Related methods	
	4.3.1	Simple OLS		4.8.1	Canonical correlation analysis
	4.3.2	Multiple OLS		4.8.2	Ridge and Lasso regression
	4.3.3	Multivariate OLS		4.8.3	Nonlinear regression
	4.4	Robust regression	4.9	Examples	
	4.4.1	Overview		4.9.1	GC retention indices of polycyclic aromatic compounds
	4.4.2	Regression diagnostics		4.9.2	Cereal data
	4.4.3	Practical hints	4.10	Summary	
	4.5	Variable selection	5	Classification	
	4.5.1	Overview		5.1	Concepts
	4.5.2	Univariate and bivariate selection methods		5.2	Linear classification methods
	4.5.3	Stepwise selection methods		5.2.1	Linear discriminant analysis
	4.5.4	Best-subset regression		5.2.2	Linear regression for discriminant analysis

	5.2.3	Logistic regression
5.3		Kernel and prototype methods
	5.3.1	SIMCA
	5.3.2	Gaussian mixture models
	5.3.3	<i>k</i> NN - classification
5.4		Classification trees
5.5		Artificial neural networks
5.6		Support vector machine
5.7		Evaluation
	5.7.1	Principles and misclassification error
	5.7.2	Predictive ability
	5.7.3	Confidence in classification answers
5.8		Examples
	5.8.1	Origin of glass samples
	5.8.2	Recognition of chemical substructures from mass spectra
5.9		Summary
6		Cluster analysis
	6.1	Concepts
	6.2	Distance and similarity measures
	6.3	Partitioning methods
	6.4	Hierarchical clustering methods
	6.5	Fuzzy clustering
	6.6	Model-based clustering
	6.7	Cluster validity and clustering tendency measures
	6.8	Examples
	6.8.1	Chemotaxonomy of plants
	6.8.2	Glass samples
	6.9	Summary

7		Preprocessing
	7.1	Concepts
	7.2	Smoothing and differentiation
	7.3	Multiplicative signal correction
	7.4	Mass spectral features

Appendix 1 Symbols and abbreviations

Appendix 2 Matrix algebra

A.2.1	Definitions
A.2.2	Addition and subtraction of matrices
A.2.3	Multiplication of vectors
A.2.4	Multiplication of matrices
A.2.5	Matrix inversion
A.2.6	Eigenvectors
A.2.7	Singular value decomposition

Appendix 3 Introduction to \mathbb{R}

A.3.1	General information on \mathbb{R}
A.3.2	Installing \mathbb{R}
A.3.3	Starting \mathbb{R}
A.3.4	Working directory
A.3.5	Loading and saving data
A.3.6	Important \mathbb{R} functions
A.3.7	Operators and basic functions
A.3.8	Data types
A.3.9	Data structures
A.3.10	Selection and extraction from data objects
A.3.11	Generating and saving graphics

Index