

Aims | Contents (*Book chapter*)

Varmuza K.:

In Engel T., Gasteiger J. (eds):

Cheminformatics - Basic Concepts and Methods,

Wiley-VCH, Weinheim, Germany (2018) p. 399-437. ISBN: 978-3-527-33109-3.

Data Analysis and Data Handling (QSPR/QSAR) - Methods for Multivariate Data Analysis.

Aims

This introduction into the use of multivariate data analysis in chemistry has manifold aims:

- It tries to help getting a basic understanding of the most used methods.
- It provides fundamental mathematical concepts for the interested reader.
- It includes some method codes for the **R** programming environment for practical numerical experiments.
- It includes examples with the data and the software provided. The examples are considered to support a better understanding and use of the methods.

Contents

11 Data Analysis and Data Handling (QSPR/QSAR) 397

11.1 Methods for Multivariate Data Analysis 399

Kurt Varmuza

11.1.1 Introduction into Multivariate Data Analysis 399

11.1.1.1 Aims 399

11.1.1.2 Notation and Symbols 400

11.1.2 Basics of Statistical Data Evaluation 401

11.1.2.1 Data Distribution, Central Value, and Spread 401

11.1.2.2 Correlation 404

11.1.2.3 Discrimination 405

11.1.3 Multivariate Data 406

11.1.3.1 Overview 406

11.1.3.2 Preprocessing 407

11.1.3.3 Distances and Similarities 408

11.1.3.4 Linear Latent Variables 410

11.1.4 Evaluation of Empirical Models 412

11.1.4.1 Overview 412

11.1.4.2 Optimum Model Complexity 412

11.1.4.3 Performance Criteria for Calibration Models 413

11.1.4.4 Performance Criteria for Classification Models 414

11.1.4.5 Cross-Validation 415

11.1.4.6 Bootstrap 416

11.1.5 Exploration: Analyzing the Independent Variables 417

11.1.5.1 Overview 417

11.1.5.2 Principal Component Analysis (PCA) 417

11.1.5.3 Nonlinear Mapping 419

11.1.5.4 Cluster Analysis 419

11.1.5.5 Example: Exploratory Data Analysis of Mass Spectra from Meteorite Samples 421

11.1.6 Calibration: Building a Quantitative Model 423

11.1.6.1 Overview 423

11.1.6.2 Ordinary Least Squares (OLS) Regression 424

11.1.6.3 Principal Component Regression (PCR) 424

11.1.6.4 Partial Least Squares (PLS) Regression 425

11.1.6.5 Variable Selection 426

11.1.6.6 Example: Prediction of Gas Chromatographic Retention Indices for Polycyclic Aromatic Hydrocarbons 427

11.1.7 Classification: Discriminating Samples 428

11.1.7.1 Overview 428

11.1.7.2 Linear Discriminant Analysis (LDA) 430

11.1.7.3 Discriminant Partial Least Squares (D-PLS) Analysis 430

11.1.7.4 k -Nearest Neighbor (KNN) Classification 430

11.1.7.5 Support Vector Machine (SVM) 431

11.1.7.6 Classification Trees (CART) 432

11.1.7.7 Example: Classification of Meteorite Samples Using Mass Spectral Data 432

Acknowledgements 434

Selected Reading 435

References 435