

# Contents

Preface.....	ix
Acknowledgments.....	xi
Authors.....	xiii
<b>Chapter 1</b> Introduction.....	1
1.1 Chemoinformatics–Chemometrics–Statistics.....	1
1.2 This Book.....	3
1.3 Historical Remarks about Chemometrics .....	4
1.4 Bibliography.....	6
1.5 Starting Examples .....	8
1.5.1 Univariate versus Bivariate Classification .....	8
1.5.2 Nitrogen Content of Cereals Computed from NIR Data.....	9
1.5.3 Elemental Composition of Archaeological Glasses.....	10
1.6 Univariate Statistics—A Reminder.....	12
1.6.1 Empirical Distributions .....	12
1.6.2 Theoretical Distributions.....	16
1.6.3 Central Value .....	19
1.6.4 Spread .....	20
1.6.5 Statistical Tests .....	22
References.....	25
<b>Chapter 2</b> Multivariate Data.....	31
2.1 Definitions.....	31
2.2 Basic Preprocessing .....	33
2.2.1 Data Transformation .....	34
2.2.2 Centering and Scaling.....	35
2.2.3 Normalization.....	36
2.2.4 Transformations for Compositional Data .....	37
2.3 Covariance and Correlation .....	38
2.3.1 Overview.....	38
2.3.2 Estimating Covariance and Correlation.....	40
2.4 Distances and Similarities.....	44
2.5 Multivariate Outlier Identification .....	47
2.6 Linear Latent Variables.....	50
2.6.1 Overview.....	50
2.6.2 Projection and Mapping.....	51
2.6.3 Example .....	53
2.7 Summary .....	56
References.....	58



<b>Chapter 3</b>	<b>Principal Component Analysis</b> .....	59
3.1	Concepts .....	59
3.2	Number of PCA Components .....	63
3.3	Centering and Scaling .....	64
3.4	Outliers and Data Distribution .....	66
3.5	Robust PCA .....	67
3.6	Algorithms for PCA .....	69
	3.6.1 Mathematics of PCA .....	69
	3.6.2 Jacobi Rotation .....	71
	3.6.3 Singular Value Decomposition .....	72
	3.6.4 NIPALS .....	73
3.7	Evaluation and Diagnostics .....	75
	3.7.1 Cross Validation for Determination of the Number of Principal Components .....	75
	3.7.2 Explained Variance for Each Variable .....	77
	3.7.3 Diagnostic Plots .....	78
3.8	Complementary Methods for Exploratory Data Analysis .....	81
	3.8.1 Factor Analysis .....	82
	3.8.2 Cluster Analysis and Dendrogram .....	82
	3.8.3 Kohonen Mapping .....	84
	3.8.4 Sammon's Nonlinear Mapping .....	87
	3.8.5 Multiway PCA .....	89
3.9	Examples .....	91
	3.9.1 Tissue Samples from Human Mummies and Fatty Acid Concentrations .....	91
	3.9.2 Polycyclic Aromatic Hydrocarbons in Aerosol .....	96
3.10	Summary .....	99
	References .....	101
<b>Chapter 4</b>	<b>Calibration</b> .....	103
4.1	Concepts .....	103
4.2	Performance of Regression Models .....	108
	4.2.1 Overview .....	108
	4.2.2 Overfitting and Underfitting .....	110
	4.2.3 Performance Criteria .....	112
	4.2.4 Criteria for Models with Different Numbers of Variables .....	114
	4.2.5 Cross Validation .....	115
	4.2.6 Bootstrap .....	118
4.3	Ordinary Least-Squares Regression .....	119
	4.3.1 Simple OLS .....	119
	4.3.2 Multiple OLS .....	124
	4.3.2.1 Confidence Intervals and Statistical Tests in OLS .....	126
	4.3.2.2 Hat Matrix and Full Cross Validation in OLS .....	129
	4.3.3 Multivariate OLS .....	129



4.4	Robust Regression .....	131
4.4.1	Overview .....	131
4.4.2	Regression Diagnostics .....	133
4.4.3	Practical Hints .....	137
4.5	Variable Selection .....	137
4.5.1	Overview .....	137
4.5.2	Univariate and Bivariate Selection Methods .....	139
4.5.3	Stepwise Selection Methods .....	140
4.5.4	Best-Subset Regression .....	141
4.5.5	Variable Selection Based on PCA or PLS Models .....	143
4.5.6	Genetic Algorithms .....	143
4.5.7	Cluster Analysis of Variables .....	146
4.5.8	Example .....	146
4.6	Principal Component Regression .....	148
4.6.1	Overview .....	148
4.6.2	Number of PCA Components .....	150
4.7	Partial Least-Squares Regression .....	150
4.7.1	Overview .....	150
4.7.2	Mathematical Aspects .....	154
4.7.3	Kernel Algorithm for PLS .....	157
4.7.4	NIPALS Algorithm for PLS .....	158
4.7.5	SIMPLS Algorithm for PLS .....	160
4.7.6	Other Algorithms for PLS .....	161
4.7.7	Robust PLS .....	162
4.8	Related Methods .....	163
4.8.1	Canonical Correlation Analysis .....	163
4.8.2	Ridge and Lasso Regression .....	166
4.8.3	Nonlinear Regression .....	168
4.8.3.1	Basis Expansions .....	168
4.8.3.2	Kernel Methods .....	169
4.8.3.3	Regression Trees .....	170
4.8.3.4	Artificial Neural Networks .....	171
4.9	Examples .....	172
4.9.1	GC Retention Indices of Polycyclic Aromatic Compounds .....	172
4.9.1.1	Principal Component Regression .....	173
4.9.1.2	Partial Least-Squares Regression .....	177
4.9.1.3	Robust PLS .....	178
4.9.1.4	Ridge Regression .....	179
4.9.1.5	Lasso Regression .....	181
4.9.1.6	Stepwise Regression .....	182
4.9.1.7	Summary .....	184
4.9.2	Cereal Data .....	185
4.10	Summary .....	188
	References .....	190



<b>Chapter 5</b>	<b>Classification</b> .....	195
5.1	Concepts.....	195
5.2	Linear Classification Methods .....	197
5.2.1	Linear Discriminant Analysis .....	197
5.2.1.1	Bayes Discriminant Analysis .....	197
5.2.1.2	Fisher Discriminant Analysis.....	200
5.2.1.3	Example.....	204
5.2.2	Linear Regression for Discriminant Analysis.....	205
5.2.2.1	Binary Classification .....	205
5.2.2.2	Multicategory Classification with OLS.....	206
5.2.2.3	Multicategory Classification with PLS .....	207
5.2.3	Logistic Regression.....	207
5.3	Kernel and Prototype Methods .....	209
5.3.1	SIMCA.....	209
5.3.2	Gaussian Mixture Models.....	212
5.3.3	$k$ -NN Classification.....	214
5.4	Classification Trees.....	217
5.5	Artificial Neural Networks.....	221
5.6	Support Vector Machine.....	223
5.7	Evaluation .....	228
5.7.1	Principles and Misclassification Error .....	228
5.7.2	Predictive Ability .....	229
5.7.3	Confidence in Classification Answers .....	230
5.8	Examples.....	231
5.8.1	Origin of Glass Samples.....	231
5.8.1.1	Linear Discriminant Analysis.....	231
5.8.1.2	Logistic Regression.....	233
5.8.1.3	Gaussian Mixture Models .....	234
5.8.1.4	$k$ -NN Methods.....	235
5.8.1.5	Classification Trees .....	236
5.8.1.6	Artificial Neural Networks.....	237
5.8.1.7	Support Vector Machines.....	238
5.8.1.8	Overall Comparison .....	238
5.8.2	Recognition of Chemical Substructures from Mass Spectra.....	240
5.9	Summary .....	246
	References.....	247
<b>Chapter 6</b>	<b>Cluster Analysis</b> .....	251
6.1	Concepts.....	251
6.2	Distance and Similarity Measures .....	254
6.3	Partitioning Methods.....	260
6.4	Hierarchical Clustering Methods .....	263
6.5	Fuzzy Clustering .....	266
6.6	Model-Based Clustering .....	267



6.7	Cluster Validity and Clustering Tendency Measures .....	270
6.8	Examples .....	272
6.8.1	Chemotaxonomy of Plants .....	272
6.8.2	Glass Samples .....	278
6.9	Summary .....	279
	References .....	281
<b>Chapter 7</b>	<b>Preprocessing .....</b>	<b>283</b>
7.1	Concepts .....	283
7.2	Smoothing and Differentiation .....	283
7.3	Multiplicative Signal Correction .....	284
7.4	Mass Spectral Features .....	287
7.4.1	Logarithmic Intensity Ratios .....	288
7.4.2	Averaged Intensities of Mass Intervals .....	288
7.4.3	Intensities Normalized to Local Intensity Sum .....	288
7.4.4	Modulo-14 Summation .....	289
7.4.5	Autocorrelation .....	289
7.4.6	Spectra Type .....	289
7.4.7	Example .....	289
	References .....	291
<b>Appendix 1</b>	<b>Symbols and Abbreviations .....</b>	<b>293</b>
<b>Appendix 2</b>	<b>Matrix Algebra .....</b>	<b>297</b>
A.2.1	Definitions .....	297
A.2.2	Addition and Subtraction of Matrices .....	298
A.2.3	Multiplication of Vectors .....	298
A.2.4	Multiplication of Matrices .....	299
A.2.5	Matrix Inversion .....	300
A.2.6	Eigenvectors .....	301
A.2.7	Singular Value Decomposition .....	302
	References .....	303
<b>Appendix 3</b>	<b>Introduction to <math>\mathbb{R}</math> .....</b>	<b>305</b>
A.3.1	General Information on $\mathbb{R}$ .....	305
A.3.2	Installing $\mathbb{R}$ .....	305
A.3.3	Starting $\mathbb{R}$ .....	305
A.3.4	Working Directory .....	306
A.3.5	Loading and Saving Data .....	306
A.3.6	Important $\mathbb{R}$ Functions .....	306
A.3.7	Operators and Basic Functions .....	307
	Mathematical and Logical Operators, Comparison .....	307
	Special Elements .....	308



	Mathematical Functions .....	308
	Matrix Manipulation.....	308
	Statistical Functions.....	308
A.3.8	Data Types .....	309
	Missing Values .....	309
A.3.9	Data Structures .....	309
A.3.10	Selection and Extraction from Data Objects.....	310
	Examples for Creating Vectors .....	310
	Examples for Selecting Elements from a Vector or Factor .....	310
	Examples for Selecting Elements from a Matrix, Array, or Data Frame.....	310
	Examples for Selecting Elements from a List.....	310
A.3.11	Generating and Saving Graphics .....	311
	Functions Relevant for Graphics.....	311
	Relevant Plot Parameters.....	311
	Statistical Graphics .....	311
	Saving Graphic Output.....	311
References	.....	312
<b>Index</b> .....		<b>313</b>