

Contents

| | |
|---|----------|
| Preface | xiii |
| 1 Introduction | 1 |
| 1.1 Development of Chemometrics | 1 |
| 1.1.1 Early Developments | 1 |
| 1.1.2 1980s and the Borderlines between Other Disciplines | 1 |
| 1.1.3 1990s and Problems of Intermediate Complexity | 2 |
| 1.1.4 Current Developments in Complex Problem Solving | 2 |
| 1.2 Application Areas | 3 |
| 1.3 How to Use this Book | 4 |
| 1.4 Literature and Other Sources of Information | 5 |
| References | 7 |
| 2 Experimental Design | 9 |
| 2.1 Why Design Experiments in Chemistry? | 9 |
| 2.2 Degrees of Freedom and Sources of Error | 12 |
| 2.3 Analysis of Variance and Interpretation of Errors | 16 |
| 2.4 Matrices, Vectors and the Pseudoinverse | 20 |
| 2.5 Design Matrices | 22 |
| 2.6 Factorial Designs | 25 |
| 2.6.1 Extending the Number of Factors | 28 |
| 2.6.2 Extending the Number of Levels | 28 |
| 2.7 An Example of a Factorial Design | 29 |
| 2.8 Fractional Factorial Designs | 32 |
| 2.9 Plackett–Burman and Taguchi Designs | 35 |
| 2.10 The Application of a Plackett–Burman Design to the Screening of Factors Influencing a Chemical Reaction | 37 |
| 2.11 Central Composite Designs | 39 |
| 2.12 Mixture Designs | 44 |
| 2.12.1 Simplex Centroid Designs | 45 |
| 2.12.2 Simplex Lattice Designs | 47 |
| 2.12.3 Constrained Mixture Designs | 47 |

| | | |
|----------|--|-----------|
| 2.13 | A Four Component Mixture Design Used to Study Blending of Olive Oils | 49 |
| 2.14 | Simplex Optimization | 51 |
| 2.15 | Leverage and Confidence in Models | 53 |
| 2.16 | Designs for Multivariate Calibration | 58 |
| | References | 62 |
| 3 | Statistical Concepts | 63 |
| 3.1 | Statistics for Chemists | 63 |
| 3.2 | Errors | 64 |
| 3.2.1 | Sampling Errors | 65 |
| 3.2.2 | Sample Preparation Errors | 66 |
| 3.2.3 | Instrumental Noise | 67 |
| 3.2.4 | Sources of Error | 67 |
| 3.3 | Describing Data | 67 |
| 3.3.1 | Descriptive Statistics | 68 |
| 3.3.2 | Graphical Presentation | 69 |
| 3.3.3 | Covariance and Correlation Coefficient | 72 |
| 3.4 | The Normal Distribution | 73 |
| 3.4.1 | Error Distributions | 73 |
| 3.4.2 | Normal Distribution Functions and Tables | 74 |
| 3.4.3 | Applications | 75 |
| 3.5 | Is a Distribution Normal? | 76 |
| 3.5.1 | Cumulative Frequency | 76 |
| 3.5.2 | Kolmogorov–Smirnov Test | 78 |
| 3.5.3 | Consequences | 79 |
| 3.6 | Hypothesis Tests | 80 |
| 3.7 | Comparison of Means: the <i>t</i> -Test | 81 |
| 3.8 | <i>F</i> -Test for Comparison of Variances | 85 |
| 3.9 | Confidence in Linear Regression | 89 |
| 3.9.1 | Linear Calibration | 90 |
| 3.9.2 | Example | 90 |
| 3.9.3 | Confidence of Prediction of Parameters | 92 |
| 3.10 | More about Confidence | 93 |
| 3.10.1 | Confidence in the Mean | 93 |
| 3.10.2 | Confidence in the Standard Deviation | 95 |
| 3.11 | Consequences of Outliers and How to Deal with Them | 96 |
| 3.12 | Detection of Outliers | 100 |
| 3.12.1 | Normal Distributions | 100 |
| 3.12.2 | Linear Regression | 101 |
| 3.12.3 | Multivariate Calibration | 103 |
| 3.13 | Shewhart Charts | 104 |
| 3.14 | More about Control Charts | 106 |
| 3.14.1 | Cusum Chart | 106 |
| 3.14.2 | Range Chart | 108 |
| 3.14.3 | Multivariate Statistical Process Control | 108 |
| | References | 109 |

| | |
|--|------------|
| 4 Sequential Methods | 111 |
| 4.1 Sequential Data | 111 |
| 4.2 Correlograms | 112 |
| 4.2.1 Auto-correlograms | 113 |
| 4.2.2 Cross-correlograms | 115 |
| 4.2.3 Multivariate Correlograms | 115 |
| 4.3 Linear Smoothing Functions and Filters | 116 |
| 4.4 Fourier Transforms | 120 |
| 4.5 Maximum Entropy and Bayesian Methods | 124 |
| 4.5.1 Bayes' Theorem | 124 |
| 4.5.2 Maximum Entropy | 125 |
| 4.5.3 Maximum Entropy and Modelling | 126 |
| 4.6 Fourier Filters | 128 |
| 4.7 Peakshapes in Chromatography and Spectroscopy | 134 |
| 4.7.1 Principal Features | 135 |
| 4.7.2 Gaussians | 136 |
| 4.7.3 Lorentzians | 136 |
| 4.7.4 Asymmetric Peak Shapes | 137 |
| 4.7.5 Use of Peak Shape Information | 138 |
| 4.8 Derivatives in Spectroscopy and Chromatography | 138 |
| 4.9 Wavelets | 142 |
| References | 143 |
| | |
| 5 Pattern Recognition | 145 |
| 5.1 Introduction | 145 |
| 5.1.1 Exploratory Data Analysis | 145 |
| 5.1.2 Unsupervised Pattern Recognition | 146 |
| 5.1.3 Supervised Pattern Recognition | 146 |
| 5.2 Principal Components Analysis | 147 |
| 5.2.1 Basic Ideas | 147 |
| 5.2.2 Method | 150 |
| 5.3 Graphical Representation of Scores and Loadings | 154 |
| 5.3.1 Case Study 1 | 154 |
| 5.3.2 Case Study 2 | 154 |
| 5.3.3 Scores Plots | 156 |
| 5.3.4 Loadings Plots | 157 |
| 5.3.5 Extensions | 159 |
| 5.4 Comparing Multivariate Patterns | 159 |
| 5.5 Preprocessing | 160 |
| 5.6 Unsupervised Pattern Recognition: Cluster Analysis | 167 |
| 5.7 Supervised Pattern Recognition | 171 |
| 5.7.1 Modelling the Training Set | 171 |
| 5.7.2 Test Sets, Cross-validation and the Bootstrap | 172 |
| 5.7.3 Applying the Model | 174 |
| 5.8 Statistical Classification Techniques | 174 |
| 5.8.1 Univariate Classification | 175 |
| 5.8.2 Bivariate and Multivariate Discriminant Models | 175 |

| | | |
|----------|---|------------|
| 5.8.3 | SIMCA | 178 |
| 5.8.4 | Statistical Output | 182 |
| 5.9 | K Nearest Neighbour Method | 182 |
| 5.10 | How Many Components Characterize a Dataset? | 185 |
| 5.11 | Multiway Pattern Recognition | 187 |
| 5.11.1 | Tucker3 Models | 188 |
| 5.11.2 | PARAFAC | 189 |
| 5.11.3 | Unfolding | 190 |
| | References | 190 |
| 6 | Calibration | 193 |
| 6.1 | Introduction | 193 |
| 6.2 | Univariate Calibration | 195 |
| 6.2.1 | Classical Calibration | 195 |
| 6.2.2 | Inverse Calibration | 196 |
| 6.2.3 | Calibration Equations | 198 |
| 6.2.4 | Including Extra Terms | 199 |
| 6.2.5 | Graphs | 199 |
| 6.3 | Multivariate Calibration and the Spectroscopy of Mixtures | 202 |
| 6.4 | Multiple Linear Regression | 206 |
| 6.5 | Principal Components Regression | 208 |
| 6.6 | Partial Least Squares | 211 |
| 6.7 | How Good is the Calibration and What is the Most Appropriate Model? | 214 |
| 6.7.1 | Autoprediction | 214 |
| 6.7.2 | Cross-validation | 215 |
| 6.7.3 | Test Sets | 215 |
| 6.7.4 | Bootstrap | 217 |
| 6.8 | Multiway Calibration | 217 |
| 6.8.1 | Unfolding | 217 |
| 6.8.2 | Trilinear PLS1 | 218 |
| 6.8.3 | N-PLSM | 219 |
| | References | 220 |
| 7 | Coupled Chromatography | 221 |
| 7.1 | Introduction | 221 |
| 7.2 | Preparing the Data | 222 |
| 7.2.1 | Preprocessing | 222 |
| 7.2.2 | Variable Selection | 224 |
| 7.3 | Chemical Composition of Sequential Data | 228 |
| 7.4 | Univariate Purity Curves | 230 |
| 7.5 | Similarity Based Methods | 234 |
| 7.5.1 | Similarity | 234 |
| 7.5.2 | Correlation Coefficients | 234 |
| 7.5.3 | Distance Measures | 235 |
| 7.5.4 | OPA and SIMPLISMA | 236 |
| 7.6 | Evolving and Window Factor Analysis | 236 |

| | |
|---|------------|
| 7.6.1 Expanding Windows | 237 |
| 7.6.2 Fixed Sized Windows | 238 |
| 7.6.3 Variations | 239 |
| 7.7 Derivative Based Methods | 239 |
| 7.8 Deconvolution of Evolutionary Signals | 241 |
| 7.9 Noniterative Methods for Resolution | 242 |
| 7.9.1 Selectivity: Finding Pure Variables | 242 |
| 7.9.2 Multiple Linear Regression | 243 |
| 7.9.3 Principal Components Regression | 244 |
| 7.9.4 Partial Selectivity | 244 |
| 7.10 Iterative Methods for Resolution | 246 |
| 8 Equilibria, Reactions and Process Analytics | 249 |
| 8.1 The Study of Equilibria using Spectroscopy | 249 |
| 8.2 Spectroscopic Monitoring of Reactions | 252 |
| 8.2.1 Mid Infrared Spectroscopy | 253 |
| 8.2.2 Near Infrared Spectroscopy | 254 |
| 8.2.3 UV/Visible Spectroscopy | 255 |
| 8.2.4 Raman Spectroscopy | 256 |
| 8.2.5 Summary of Main Data Analysis Techniques | 256 |
| 8.3 Kinetics and Multivariate Models for the Quantitative Study of Reactions | 257 |
| 8.4 Developments in the Analysis of Reactions using On-line Spectroscopy | 261 |
| 8.4.1 Constraints and Combining Information | 261 |
| 8.4.2 Data Merging | 262 |
| 8.4.3 Three-way Analysis | 262 |
| 8.5 The Process Analytical Technology Initiative | 263 |
| 8.5.1 Multivariate Tools for Design, Data Acquisition and Analysis | 263 |
| 8.5.2 Process Analysers | 264 |
| 8.5.3 Process Control Tools | 264 |
| 8.5.4 Continuous Improvement and Knowledge Management Tools | 264 |
| References | 265 |
| 9 Improving Yields and Processes Using Experimental Designs | 267 |
| 9.1 Introduction | 267 |
| 9.2 Use of Statistical Designs for Improving the Performance of Synthetic Reactions | 269 |
| 9.3 Screening for Factors that Influence the Performance of a Reaction | 271 |
| 9.4 Optimizing the Process Variables | 275 |
| 9.5 Handling Mixture Variables using Simplex Designs | 278 |
| 9.5.1 Simplex Centroid and Lattice Designs | 278 |
| 9.5.2 Constraints | 280 |
| 9.6 More about Mixture Variables | 283 |
| 9.6.1 Ratios | 283 |
| 9.6.2 Minor Constituents | 285 |
| 9.6.3 Combining Mixture and Process Variables | 285 |
| 9.6.4 Models | 285 |

| | |
|--|------------|
| 10 Biological and Medical Applications of Chemometrics | 287 |
| 10.1 Introduction | 287 |
| 10.1.1 Genomics, Proteomics and Metabolomics | 287 |
| 10.1.2 Disease Diagnosis | 288 |
| 10.1.3 Chemical Taxonomy | 288 |
| 10.2 Taxonomy | 289 |
| 10.3 Discrimination | 291 |
| 10.3.1 Discriminant Function | 291 |
| 10.3.2 Combining Parameters | 293 |
| 10.3.3 Several Classes | 293 |
| 10.3.4 Limitations | 296 |
| 10.4 Mahalanobis Distance | 297 |
| 10.5 Bayesian Methods and Contingency Tables | 300 |
| 10.6 Support Vector Machines | 303 |
| 10.7 Discriminant Partial Least Squares | 306 |
| 10.8 Micro-organisms | 308 |
| 10.8.1 Mid Infrared Spectroscopy | 309 |
| 10.8.2 Growth Curves | 311 |
| 10.8.3 Further Measurements | 311 |
| 10.8.4 Pyrolysis Mass Spectrometry | 312 |
| 10.9 Medical Diagnosis using Spectroscopy | 313 |
| 10.10 Metabolomics using Coupled Chromatography and Nuclear Magnetic Resonance | 314 |
| 10.10.1 Coupled Chromatography | 314 |
| 10.10.2 Nuclear Magnetic Resonance | 317 |
| References | 317 |
| 11 Biological Macromolecules | 319 |
| 11.1 Introduction | 319 |
| 11.2 Sequence Alignment and Scoring Matches | 320 |
| 11.3 Sequence Similarity | 322 |
| 11.4 Tree Diagrams | 324 |
| 11.4.1 Diagrammatic Representations | 324 |
| 11.4.2 Dendrograms | 325 |
| 11.4.3 Evolutionary Theory and Cladistics | 325 |
| 11.4.4 Phylograms | 326 |
| 11.5 Phylogenetic Trees | 327 |
| References | 329 |
| 12 Multivariate Image Analysis | 331 |
| 12.1 Introduction | 331 |
| 12.2 Scaling Images | 333 |
| 12.2.1 Scaling Spectral Variables | 334 |
| 12.2.2 Scaling Spatial Variables | 334 |
| 12.2.3 Multiway Image Preprocessing | 334 |
| 12.3 Filtering and Smoothing the Image | 335 |
| 12.4 Principal Components for the Enhancement of Images | 337 |

| | | |
|-----------|--|------------|
| 12.5 | Regression of Images | 340 |
| 12.6 | Alternating Least Squares as Employed in Image Analysis | 345 |
| 12.7 | Multiway Methods In Image Analysis | 347 |
| | References | 349 |
| 13 | Food | 351 |
| 13.1 | Introduction | 351 |
| 13.1.1 | Adulteration | 351 |
| 13.1.2 | Ingredients | 352 |
| 13.1.3 | Sensory Studies | 352 |
| 13.1.4 | Product Quality | 352 |
| 13.1.5 | Image Analysis | 352 |
| 13.2 | How to Determine the Origin of a Food Product using Chromatography | 353 |
| 13.3 | Near Infrared Spectroscopy | 354 |
| 13.3.1 | Calibration | 354 |
| 13.3.2 | Classification | 355 |
| 13.3.3 | Exploratory Methods | 355 |
| 13.4 | Other Information | 356 |
| 13.4.1 | Spectroscopies | 356 |
| 13.4.2 | Chemical Composition | 356 |
| 13.4.3 | Mass Spectrometry and Pyrolysis | 357 |
| 13.5 | Sensory Analysis: Linking Composition to Properties | 357 |
| 13.5.1 | Sensory Panels | 357 |
| 13.5.2 | Principal Components Analysis | 359 |
| 13.5.3 | Advantages | 359 |
| 13.6 | Varimax Rotation | 359 |
| 13.7 | Calibrating Sensory Descriptors to Composition | 365 |
| | References | 368 |
| | Index | 369 |

