DATA HANDLING IN SCIENCE AND TECHNOLOGY — VOLUME 2

Advisory Editors: B.G.M. Vandeginste and L. Kaufman

Chemometrics: a textbook

D.L. MASSART
Farmaceutisch Instituut, Vrije Universiteit Brussel, Belgium

B.G.M. VANDEGINSTE
Laboratorium voor Analytische Chemie, Katholieke Universiteit Nijmegen, The Netherlands

S.N. DEMING
Department of Chemistry, University of Houston, Texas, U.S.A.

Y. MICHOTTE and L. KAUFMAN
Farmaceutisch Instituut, Vrije Universiteit Brussel, Belgium

ELSEVIER

Amsterdam — Oxford — New York — Tokyo 1988

CONTENTS

Introduction 1
Reference 3

Chapter 1. Chemometrics and the Analytical Process 5
1. Definition of chemometrics 5
2. The intelligent laboratory concept 5
3. Organization of the book 9
Recommended reading 10

Chapter 2. Precision and Accuracy 13
1. General discussion of concepts 13
1.1 Introduction 13
1.2 Categories of errors in analytical chemistry 13
1.3 Precision and accuracy as criteria 15
1.4 Definition and measurement of precision (repeatability, reproducibility) 15
1.5 Definitions of bias and accuracy 18
1.6 Laboratory bias 19
2. Mathematical 22
2.1 Discrete and continuous random variables 22
2.2 Probability functions. Cumulative frequency distribution functions 22
2.2.1 Continuous variables 22
2.2.2 Discrete variables 23
2.3 Mean and variance in samples and populations 23
2.4 Mean and variance after transformation of the variable x 24
2.5 Mean and variance of a linear function of random variables 24
2.6 The normal distribution 25
2.6.1 Statistical description 25
2.6.2 The standardized normal distribution 26
2.6.3 The central limit theorem 27
2.7 Other distributions 28
2.7.1 The binomial distribution 28
2.7.2 The Poisson distribution 28
2.7.3 The chi-square distribution 29
2.7.4 The t-distribution 29
2.7.5 The F-distribution 30
References 31
Recommended reading 31

Chapter 3. Evaluation of Precision and Accuracy. Comparison of Two Procedures 33
1. Introduction 33
2. Statistics 34
2.1 The elaboration of a test 34
2.2 Types of error 35
2.3 One- and two-sided tests 38
2.4 Statistical scales 39
2.4.1 The nominal scale 40
2.4.2 The ordinal scale 40
2.4.3 The interval scale 40
2.4.4 The ratio scale 40

3. Evaluation of method bias using tests on the mean (t-test) . . . 41
3.1 The one-sample case . . . 42
3.2 The two-sample case . . . 43
3.2.1 Comparison of the means of two independent samples . . . 44
3.2.2 Comparison of the means of two related (paired) samples . . . 47
4. Non-parametric tests for the comparison of methods . . . 48
4.1 The Mann–Whitney U-test (independent samples) . . . 50
4.2 The Wilcoxon matched-pairs signed-ranks test or the Wilcoxon T-test (related samples) . . . 51
5. Comparison of two methods by least-squares fitting . . . 52
6. Recovery experiments . . . 54
7. Comparison of the precision of different methods (F-test) . . . 55
8. The χ^2-test . . . 56
References . . . 57
Recommended reading . . . 57
Software . . . 57

Chapter 4. Evaluation of Sources of Variation in Data. Analysis of Variance . . . 59
1. Introduction. Definitions . . . 59
2. The one-way layout . . . 60
3. A posteriori tests . . . 65
4. The two-way layout . . . 66
5. Bartlett's test for the comparison of more than two variances . . . 70
6. Applications . . . 71
References . . . 73
Recommended books on ANOVA . . . 73
Recommended reading . . . 74

Chapter 5. Calibration . . . 75
1. Introduction . . . 75
2. The method of least squares . . . 75
3. Analysis of residuals . . . 78
3.1 Graphical methods . . . 78
3.2 Statistical methods . . . 80
4. Analysis of variance . . . 80
5. Heteroscedasticity . . . 84
5.1 Weighting . . . 84
5.1.1 Weighted least squares . . . 84
5.1.2 Determination of weighting factors . . . 85
5.2 Transformation to homoscedasticity . . . 86
6. Confidence intervals . . . 86
7. Some further developments . . . 91
References . . . 92
Recommended reading . . . 92

Chapter 6. Reliability and Drift . . . 93
1. The a posteriori approach. Trend detection . . . 94
1.1 Control chart methods for detection of drift . . . 94
1.2 The Cusum technique . . . 96
1.3 Trigg's monitoring technique . . . 97
1.4 Other statistical methods . . . 99
2. The a priori approach. Ruggedness of a method . . . 101
References . . . 105
Recommended reading . . . 106

Chapter 7. Sensitivity and Limit of Detection . . . 107
1. Introduction . . . 107
2. Sensitivity and the analytical calibration function . . . 108
3. Decision limit . . . 110
4. Detection limit . . . 112
5. Determination limit . . . 113
6. Discussion . . . 113
References . . . 114
Recommended reading . . . 114

Chapter 8. Selectivity and Specificity . . . 115
1. Interferences and matrix effects . . . 115
2. Qualitative definition of specificity and selectivity . . . 117
3. Quantitative definition of selectivity and specificity . . . 118
3.1 Propagation of error . . . 118
3.2 Quantitative definition . . . 119
4. Choosing the optimal set of wavenumbers . . . 123
References . . . 126
Recommended reading . . . 126

Chapter 9. Information . . . 127
1. Introduction . . . 127
2. An application to thin layer chromatography . . . 130
3. The information content of combined procedures . . . 132
References . . . 135
Recommended reading . . . 136

Chapter 10. Costs . . . 137
1. The cost of a determination . . . 137
2. The cost and the benefit of analytical information . . . 141
2.1 Cost–benefit analysis. Opportunity costs . . . 141
2.2 The law of marginal diminishing utility and related laws . . . 142
References . . . 146
Recommended reading . . . 146

Chapter 11. The Time Constant . . . 149
References . . . 156
Recommended reading . . . 156

Chapter 12. Signals and Data . . . 157
1. Classification of signals and data . . . 157
1.1 Deterministic signals . . . 157
1.2 Stochastic signals . . . 158
1.3 Signals with a deterministic and a stochastic part . . . 159
1.4 Univariate and multivariate data and signals . . . 159
2. Purpose of signal and data processing . . . 161
2.1 Signal enhancement by reduction of the noise level . . . 161
2.2 Signal restoration . . . 161
2.3 Characterization and modelling of the stochastic and deterministic parts of the signal . . . 161
2.3.1 Univariate data . . . 161
2.3.2 Multivariate data . . . 162
2.4 Data reduction . . . 163
Reference . . . 163

Chapter 13. Regression Methods 165
1. Introduction 165
2. Univariate regression 167
2.1 Linear regression 167
2.1.1 Parameter estimation 167
2.1.2 Confidence intervals for the parameters 172
2.2 Non-linear regression 175
2.2.1 Parameter estimation 175
2.2.2 Validation of the non-linear model 180
3. Multivariate linear regression 181
3.1 Fitting of plane surfaces 181
3.2 Multicomponent analysis by multiple linear regression 185
References 188
Recommended reading 188

Chapter 14. Correlation Methods 191
1. Introduction 191
2. Covariance and correlation 193
3. Correlation and regression 196
4. The variance–covariance matrix 197
5. Time series of random variables 201
5.1 Autocovariance and autocorrelation 202
5.2 Autocorrelograms of some characteristic processes 206
5.3 Practical examples 209
5.4 Autoregression 211
References 212
Recommended reading 212

Chapter 15. Signal Processing 215
1. The Fourier transform 217
1.1 The fast Fourier transform (FFT) 220
1.2 The inverse Fourier transform 221
1.3 Properties of the Fourier transform 222
1.3.1 Periodicity and symmetry 222
1.3.2 Distributivity 223
1.3.3 Shift 224
1.3.4 Convolution 224
2. Analytical applications 230
2.1 Deconvolution (signal restoration) 230
2.2 Signal and noise characterization 233
2.3 Filtering 236
2.4 Digitization of analytical signals 240
3. Noise reduction by direct smoothing of the data 241
3.1 Noise removal 241
3.2 Smoothing 242
3.2.1 Moving average 242
3.2.2 Exponential averaging 242
3.2.3 Polynomial smoothing with Savitzky–Golay filters 244
References 252
Recommended reading 252

Chapter 16. Response Surfaces and Models 255
1. Introduction 255

2. Response surfaces . . . 255
2.1 True response surfaces . . . 255
2.2 Measured response surfaces . . . 257
2.3 Estimated response surfaces and models . . . 258
3. Two-factor response surfaces and models . . . 261
3.1 First-order models . . . 262
3.2 First-order models with interaction . . . 264
3.3 Full second-order polynomial models . . . 266
Reference . . . 269
Recommended reading . . . 269

Chapter 17. Exploration of Response Surfaces . . . 271
1. Introduction . . . 271
2. Two-level factorial designs . . . 271
3. Qualitative factors . . . 273
4. Cautions about the chosen system of coding . . . 275
5. Quantitative factors . . . 277
6. Fractional replication . . . 281
7. Central composite designs . . . 283
8. General comments on experimental designs . . . 287
9. Exploring response surfaces . . . 287
9.1 Screening experiments . . . 287
9.2 Evolutionary operation . . . 288
9.3 Mapping the region of the optimum . . . 289
References . . . 290
Recommended reading . . . 290

Chapter 18. Optimization of Analytical Chemical Methods . . . 293
1. Introduction . . . 293
2. Sequential single-factor-at-a-time strategy . . . 293
3. Box-type EVOP . . . 296
4. Simplex EVOP . . . 297
5. The fixed-size simplex algorithm . . . 299
6. The variable-size simplex algorithm . . . 301
7. Joint use of simplex and classical experimental designs . . . 303
References . . . 304
Recommended reading . . . 304

Chapter 19. Optimization of Chromatographic Methods . . . 305
1. Introduction . . . 305
2. Multifactor designs in chromatography . . . 305
3. Semi-empirical models . . . 307
4. Multifactor chromatographic optimization . . . 310
5. Ruggedness . . . 312
6. Other approaches to multifactor chromatographic optimization . . . 313
7. Sequential simplex optimization . . . 316
References . . . 316
Recommended reading . . . 316

Chapter 20. The Multivariate Approach . . . 319
1. A multivariate look at clinical chemistry. Multivariate distributions . . . 319
2. The need for visual information . . . 320
3. Patterns, classes, and distances . . . 321

4. The classification of iron meteorites. Clustering . . . 323
5. Air pollution. Principal components . . . 325
6. The geographical origin of olive oils. Supervised pattern recognition . . . 330
7. Thyroid status. Supervised pattern recognition . . . 331
8. Optimal combination of variables. Feature selection . . . 334
9. Similarity and distance . . . 335
References . . . 337
Recommended reading . . . 337

Chapter 21. Principal Components and Factor Analysis . . . 339
1. The interpretation of principal components plots . . . 341
2. Principal components in two dimensions . . . 343
3. Principal components in m dimensions . . . 349
4. Hard modelling versus soft modelling . . . 355
5. Factor analysis . . . 358
5.1 Determination of factors . . . 358
5.2 Rotation of factors . . . 359
5.3 An analytical example . . . 362
6. Other display methods . . . 366
6.1 Correspondence factor analysis . . . 366
6.2 Non-linear mapping (multidimensional scaling) . . . 366
References . . . 369
Recommended Reading . . . 369

Chapter 22. Clustering Techniques . . . 371
1. Measures of (dis)similarity . . . 372
2. Hierarchical methods . . . 375
3. Non-hierarchical methods . . . 378
4. Some other methods . . . 380
5. Selection of a clustering method . . . 381
References . . . 383
Recommended reading . . . 383

Chapter 23. Supervised Pattern Recognition . . . 385
1. Decision rules and their validation . . . 385
2. Statistical linear discriminant analysis . . . 387
3. The linear learning machine . . . 392
4. The k-nearest neighbour method . . . 395
5. Density methods . . . 397
6. Modelling methods . . . 399
6.1 Methods based on the multivariate normal distribution . . . 400
6.2 SIMCA . . . 403
7. Feature selection . . . 407
7.1 Univariate strategy . . . 407
7.2 Multivariate strategy . . . 408
7.3 Mixed multivariate–univariate strategy . . . 408
7.4 Conclusions . . . 409
8. Recommended methods and strategy . . . 409
8.1 Pretreatment of the data . . . 411
8.2 Looking at the data . . . 411
8.3 Redefinition of classes . . . 411
8.4 Feature selection . . . 411
8.5 Selection of method . . . 411

References 412
Recommended reading 412

Chapter 24. Decisions in the Analytical Laboratory 415
1. Decisions based on analytical results 416
2. Decisions about the analytical process 417
Reference 418

Chapter 25. Operations Research 419
1. A queueing problem 420
2. A shortest path problem 427
3. Operations research models used in clustering 431
4. Digital simulation 431
References 434
Recommenced reading 434

Chapter 26. Decision Making 437
1. A decision matrix 437
2. Probability theory 439
3. The operating characteristic curve and the power function in quality control 440
4. Prior and posterior probabilities. Bayes' rule 443
5. The diagnostic value of a test 445
6. The ROC curve 449
References 451
Recommended reading 451

Chapter 27. Process Control 453
1. Derivation of a process model 455
2. Quality of a control loop 457
3. An example 463
References 466
Recommended Reading 467

Appendix 469

Index 477