Multivariate Data Analysis
6th Edition

An introduction to Multivariate Analysis, Process Analytical Technology and Quality by Design

Kim H. Esbensen
and
Brad Swarbrick

with contributions from Frank Westad, Pat Whitcombe and Mark Anderson

CAMO
Bring data to life

Contents

Preface xv

Chapter 1. Introduction to multivariate analysis 1

1.1 The world is multivariate 1

1.2 Indirect observations and correlations 2

1.3 Data must carry useful information 2

1.4 Variance, covariance and correlation 3

1.5 Causality vs correlation 6

1.6 Hidden data structures—correlations again 6

1.7 Multivariate data analysis vs multivariate statistics 8

1.8 Main objectives of multivariate data analysis 8

1.8.1 Data description (exploratory data structure modelling) 9

1.8.2 Discrimination and classification 9

1.8.3 Regression and prediction 10

1.9 Multivariate techniques as geometric projections 10

1.9.1 Geometry, mathematics, algorithms 11

1.10 The grand overview in multivariate data analysis 11

1.11 References 12

Chapter 2: A review of some fundamental statistics 13

2.1 Terminology 13

2.2 Definitions of some important measurements and concepts 14

2.2.1 The mean 15

2.2.2 The median 16

2.2.3 The mode 17

2.2.4 Variance and standard deviation 17

2.3 Samples and representative sampling 18

2.3.1 An example from the pharmaceutical industry 19

2.4 The normal distribution and its properties 20

2.4.1 Graphical representations 20

2.5 Hypothesis testing....26
2.5.1 Significance, risk and power.... 26
2.5.2 Defining an appropriate risk level.... 28
2.5.3 A general guideline for applying formal statistical tests.... 30
2.5.4 A Test for Equivalence of Variances: The *F*-test.... 35
2.5.5 Tests for equivalence of means.... 38
2.6 An introduction to time series and control charts....45
2.7 Joint confidence intervals and the need for multivariate analysis.........48
2.8 Chapter summary....50
2.9 References....52

Chapter 3: Theory of Sampling (TOS) 53

3.1 Chapter overview....54
3.2 Heterogeneity....54
3.2.1 Constitutional heterogeneity (CH).... 55
3.2.2 Distributional heterogeneity (DH).... 55
3.3 Sampling error vs practical sampling....57
3.4 Total Sampling Error (TSE)—Fundamental Sampling Principle (FSP) ...58
3.5 Sampling Unit Operations (SUO)....59
3.6 Replication experiment—quantifying sampling errors....61
3.7 TOS in relation to multivariate data analysis....62
3.8 Process sampling—variographic analysis....63
3.8.1 Appendix A. Terms and definitions used in the TOS literature.... 65
3.9 References....68

Chapter 4: Fundamentals of principal component analysis (PCA) 69

4.1 Representing data as a matrix....69
4.2 The variable space—plotting objects in *p*-dimensions....70
4.2.1 Plotting data in 1-d and 2d space.... 70
4.2.2 The variable space and dimensions.... 70
4.2.3 Visualisation in 3-D (or more).... 70
4.3 Plotting objects in variable space....71
4.4 Example—plotting raw data (beverage)....71
4.4.1 Purpose.... 71
4.4.2 Data set.... 71

4.5 The first principal component ... 73
4.5.1 Maximum variance directions ... 73
4.5.2 The first principal component as a least squares fit ... 74
4.6 Extension to higher-order principal components ... 75
4.7 Principal component models—scores and loadings ... 76
4.7.1 Maximum number of principal components ... 76
4.7.2 PC model centre ... 77
4.7.3 Introducing loadings—relations between X and PCs ... 77
4.7.4 Scores—coordinates in PC space ... 78
4.7.5 Object residuals ... 78
4.8 Objectives of PCA ... 79
4.9 Score plot–object relationships ... 80
4.9.1 Interpreting score plots ... 80
4.9.2 Choice of score plots ... 82
4.10 The loading plot–variable relationships ... 83
4.10.1 Correlation loadings ... 84
4.10.2 Comparison of scores and loading plots ... 86
4.10.3 The 1-dimensional loading plot ... 87
4.11 Example: city temperatures in europe ... 89
4.11.1 Introduction ... 89
4.11.2 Plotting data and deciding on the validation scheme ... 89
4.11.3 PCA results and interpretation ... 90
4.12 Principal component models ... 93
4.12.1 The PC model ... 93
4.12.2 Centring ... 93
4.12.3 Step by step calculation of PCs ... 94
4.12.4 A preliminary comment on the algorithm: NIPALS ... 94
4.12.5 Residuals—the **E**-matrix ... 95
4.12.6 Residual variance ... 95
4.12.7 Object residuals ... 96
4.12.8 The total squared object residual ... 96
4.12.9 Explained/residual variance plots ... 96
4.12.10 How many PCs to use? ... 97
4.12.11 A note on the number of PCs ... 98
4.12.12 A doubtful case—using external evidence ... 98
4.12.13 Variable residuals ... 99
4.12.14 More about variances—modelling error variance ... 99

4.13 Example: interpreting a PCA model (peas) 99
4.13.1 Purpose 100
4.13.2 Data set 100
4.13.3 Tasks 100
4.13.4 How to do it 100
4.13.5 Summary 101
4.14 PCA modelling—the NIPALS algorithm 102
4.15 Chapter summary 103
4.16 References 104

Chapter 5: Preprocessing 106
5.1 Introduction 106
5.2 Preprocessing of discrete data 106
5.2.1 Variable weighting and scaling 106
5.2.2 Logarithm transformation 108
5.2.3 Averaging 108
5.3 Preprocessing of spectroscopic data 109
5.3.1 Spectroscopic transformations 110
5.3.2 Smoothing 112
5.3.3 Normalisation 113
5.3.4 Baseline correction 114
5.3.5 Derivatives 116
5.3.6 Correcting multiplicative effects in spectra 122
5.3.7 Other general preprocessing methods 125
5.4 Practical aspects of preprocessing 127
5.4.1 Scatter effects plot 129
5.4.2 Detailed example: preprocessing gluten–starch mixtures 130
5.5 Chapter summary 133
5.6 References 134

6. Principal Component Analysis (PCA)—in practice 135
6.1 The PCA overview 135
6.2 PCA—Step by Step 136
6.3 Interpretation of PCA models 138
6.3.1 Interpretation of score plots—look for patterns 138
6.3.2 Summary—interpretation of score plots 140
6.3.3 Interpretation of loading plots—look for important variables 140

6.4 Example: alcohol in water analysis 141

6.5 PCA—what can go wrong? 144

6.5.1 Is there any information in the data set? 144

6.5.2 Too few PCs are used in the model 145

6.5.3 Too many PCs are used in the model 145

6.5.4 Outliers which are truly due to erroneous data were not removed 145

6.5.5 Outliers that contain important information were removed 145

6.5.6 The score plots were not explored sufficiently 145

6.5.7 Loadings were interpreted with the wrong number of PCs 145

6.5.8 Too much reliance on the standard diagnostics in the computer program without thinking for yourself 145

6.5.9 The "wrong" data preprocessing was used 145

6.6 Outliers 146

6.6.1 Hotelling's T^2 statistic 147

6.6.2 Leverage 147

6.6.3 Mahalanobis distance 148

6.6.4 Influence plots 148

6.7 Validation score plot and PCA projection 149

6.7.1 Multivariate projection 150

6.7.2 Validation scores 150

6.8 Exercise—detecting outliers (Troodos) 152

6.8.1 Purpose 152

6.8.2 Data set 152

6.8.3 Analysis 153

6.8.4 Summary 156

6.9 Summary: PCA in practice 156

6.10 References 157

7. Multivariate calibration 158

7.1 Multivariate modelling (**X**, **Y**): the calibration stage 158

7.2 Multivariate modelling (**X**, **Y**): the prediction stage 159

7.3 Calibration set requirements (training set) 160

7.4 Introduction to validation 161

7.4.1 Test set validation 161

7.4.2 Other validation methods 162

7.4.3 Modelling error 162

7.5 Number of components/factors (model dimensionality) 163

7.5.1 Minimising the prediction error 163

7.6 Univariate regression (**y**|**x**) and MLR 164

7.6.1 Univariate regression (**y**|**x**) 164

7.6.2 Multiple linear regression, MLR 165

7.7 Collinearity 166

7.8 PCR—Principal component regression....166

7.8.1 PCA scores in MLR 166

7.8.2 Are all the possible PCs needed? 167

7.8.3 Example: prediction of multiple components in an alcohol mixture 168

7.8.4 Weaknesses of PCR 170

7.9 PLS-regression (PLSR) 171

7.9.1 PLSR—a powerful alternative to PCR 171

7.9.2 PLSR (**X**, **Y**): initial comparison with PCA(**X**), PCA(**Y**) 172

7.9.3 PLS—NIPALS algorithm 173

7.9.4 PLSR with one or more **Y**-variables 175

7.9.5 Interpretation of PLS models 176

7.9.6 Loadings (**p**) and loading weights (**w**) 176

7.9.7 The PLS1 NIPALS algorithm 177

7.10 Example—interpretation of PLS1 (octane in gasoline) part 1: model development 178

7.10.1 Purpose 178

7.10.2 Data set 178

7.10.3 Tasks 178

7.10.4 Initial data considerations 178

7.10.5 Always perform an initial PCA 181

7.10.6 Regression analysis 182

7.10.7 Assessment of loadings vs loading weights 182

7.10.8 Assessment of regression coefficients 183

7.10.9 Always use loading weights for model building and understanding.... 184

7.10.10 Predicted vs reference plot 185

7.10.11 Regression analysis of octane (Part 1) summary 186

7.10.12 A short discourse on model diagnostics 187

7.10.13 Residuals in X 187

7.10.14 *Q*-residuals 188

7.10.15 *F*-residuals 188

7.10.16 Hotelling's T^2 statistic 188

7.10.17 Influence plots for regression models 189

7.10.18 Always check the raw data! 189

7.10.19 Which objects should be removed? 189

7.10.20 Residuals in Y 190

7.11 Error measures 192
7.11.1 Calculating the SEL for a reference method 193
7.11.2 Further estimates of model precision 193
7.11.3 **X–Y** relation outlier plots (*T* vs *U* scores) 194
7.11.4 Example—interpretation of PLS1 (octane in gasoline) Part 2: advanced interpretations 195
7.11.5 Sample elimination 195
7.11.6 Variable elimination 196
7.11.7 **X–Y** relationship outlier plot 198
7.12 Prediction using multivariate models 199
7.12.1 Projected scores 202
7.12.2 Prediction influence plots 202
7.12.3 **Y**-deviation 203
7.12.4 Inlier statistic 203
7.12.5 Example—interpretation of PLS1 (octane in gasoline) Part 3: prediction 203
7.13 Uncertainty estimates, significance and stability—Martens' uncertainty test 205
7.13.1 Uncertainty estimates in regression coefficients, ***b*** ***206***
7.13.2 Rotation of perturbed models 206
7.13.3 Variable selection 206
7.13.4 Model stability 207
7.13.5 An example using data from paper manufacturing 207
7.13.6 Example—gluten in starch calibration 207
7.13.7 Raw data model 209
7.13.8 MSC data model 210
7.13.9 EMSC data model 210
7.13.10 mEMSC data model 211
7.13.11 Comparison of results 211
7.14 PLSR and PCR multivariate calibration—in practice 212
7.14.1 What is a "good" or "bad" model? 213
7.14.2 Signs of unsatisfactory data models—a useful checklist 214
7.14.3 Possible reasons for bad modelling or validation results 215
7.15 Chapter summary 216
7.16 References 217
8. Principles of Proper Validation (PPV) 218
8.1 Introduction 218
8.2 The Principles of Validation: overview 219
8.3 Data quality—data representativity 220

8.4 Validation objectives 220

8.4.1 Test set validation—a necessary and sufficient paradigm 221

8.4.2 Validation in data analysis and chemometrics 222

8.5 Fallacies and abuse of the central limit theorem 222

8.6 Systematics of cross-validation 222

8.7 Data structure display via *t*–*u* plots 223

8.8 Multiple validation approaches 227

8.9 Verdict on training set splitting and many other myths....................227

8.10 Cross-validation does have a role—category and model comparisons... 232

8.11 Cross-validation vs test set validation in practice....................234

8.12 Visualisation of validation is everything 234

8.13 Final remark on several test sets 235

8.14 Conclusions....................236

8.15 References 237

9. Replication—replicates—but of what? 239

9.1 Introduction....................239

9.2 Understanding uncertainty 241

9.3 The Replication Experiment (RE)....................242

9.4 RE consequences for validation 245

9.5 Replication applied to analytical method development....................245

9.6 Analytical vs sampling bias....................247

9.7 References 249

10. An introduction to multivariate classification 251

10.1 Supervised or unsupervised, that is the question!....................251

10.2 Principles of unsupervised classification and clustering 251

10.2.1 *k*-Means clustering.......... 252

10.3 Principles of supervised classification 259

10.4 Graphical interpretation of classification results....................264

10.4.1 The Coomans' plot.......... 264

10.5 Partial least squares discriminant analysis (PLS-DA) 272

10.5.1 Multivariate classification using class differences, PLS-DA.......... 272

10.6 Linear Discriminant Analysis (LDA) 275

10.7 Support vector machine classification....277

10.8 Advantages of SIMCA over traditional methods and new methods...280

10.9 Application of supervised classification methods to authentication of vegetable oils using FTIR....280

10.9.1 Data visualisation and pre-processing 280

10.9.2 Exploratory data analysis 281

10.9.3 Developing a SIMCA library and application to a test set 282

10.9.4 SIMCA model diagnostics 283

10.9.5 Developing a PLS-DA method and application to a test set.... 284

10.9.6 Developing a PCA-LDA method and application to a test set.... 285

10.9.7 Developing a SVMC method and application to a test set 288

10.9.8 Conclusions from the Vegetable Oil classification.... 288

10.10 Chapter summary 290

10.11 References 292

Chapter 11. Introduction to Design of Experiment (DoE) Methodology 293

11.1 Experimental design 293

11.1.1 Why is experimental design useful?.... 293

11.1.2 The *ad hoc* approach.... 293

11.1.3 The traditional approach—vary one variable at a time.... 294

11.1.4 The alternative approach 295

11.2 Experimental design in practice....296

11.2.1 Define stage 296

11.2.2 Design stage 296

11.2.3 Analyse stage 297

11.2.4 Improve stage.... 297

11.2.5 The concept of factorial designs 297

11.2.6 Full factorial designs.... 297

11.2.7 Naming convention 299

11.2.8 Calculating effects when there are many experiments 300

11.2.9 The concept of fractional factorial designs.... 302

11.2.10 Confounding.... 303

11.2.11 Types of variables encountered in DoE.... 305

11.2.12 Ranges of variation for experimental factors.... 307

11.2.13 Replicates.... 308

11.2.14 Randomisation.... 308

11.2.15 Blocking in designed experiments.... 309

11.2.16 Types of experimental design 309
11.2.17 Which optimisation design to choose in practice 315
11.2.18 Important effects 316
11.2.19 Hierarchy of effects 318
11.2.20 Model significance 318
11.2.21 Total sum of squares (SS_{total}) 319
11.2.22 Sum of squares regression (SS_{Reg}) 320
11.2.23 Residual sum of squares (SS_{Error}) 320
11.2.24 Model degrees of freedom (ν) 320
11.2.25 Example: building the ANOVA table for a 2^3 full factorial design 322
11.2.26 Supplementary statistics 323
11.2.27 Pure error and lack of fit assessment 330
11.2.28 Graphical tools used for assessing designed experiments 333
11.2.29 Model interpretation plots 336
11.2.30 The chemical process as a fractional factorial design 339
11.2.31 An introduction to constrained designs 352

11.3 Chapter summary 381

11.4 References 385

Chapter 12. Factor rotation and multivariate curve resolution—introduction to multivariate data analysis, tier II 387

12.1 Simple structure 387

12.2 PCA rotation 387

12.3 Orthogonal rotation methods 389

12.3.1 Varimax rotation 389
12.3.2 Quartimax rotation 389
12.3.3 Equimax rotation 390
12.3.4 Parsimax rotation 390

12.4 Interpretation of rotated PCA results 390

12.4.1 PCA rotation applied to NIR data of fish samples 390

12.5 An introduction to multivariate curve resolution (MCR) 394

12.5.1 What is multivariate curve resolution? 394
12.5.2 How multivariate curve resolution works 395
12.5.3 Data types suitable for MCR 395

12.6 Constraints in MCR 396

12.6.1 Non-negativity constraints 397
12.6.2 Uni-modality constraints 397
12.6.3 Closure constraints 398
12.6.4 Other constraints 398

12.6.5 Ambiguities and constraints in MCR 400

12.7 Algorithms used in multivariate curve resolution 401

12.7.1 Evolving factor analysis (EFA) 401

12.7.2 Multivariate curve resolution–alternating least squares (MCR–ALS) 401

12.7.3 Initial estimates for MCR-ALS 403

12.7.4 Computational parameters of MCR–ALS 403

12.7.5 Tuning the sensitivity of the analysis to pure components 404

12.8 Main results of MCR 404

12.8.1 Residuals 404

12.8.2 Estimated concentrations 405

12.8.3 Estimated spectra 405

12.8.4 Practical use of estimated concentrations and spectra and quality checks 405

12.8.5 Outliers and noisy variables in MCR 405

12.9 MCR applied to fat analysis of fish 406

12.10 Chapter summary 409

12.11 References 410

Chapter 13. Process analytical technology (PAT) and its role in the quality by design (QbD) initiative 413

13.1 Introduction 413

13.2 The Quality by Design (QbD) initiative 414

13.2.1 The International Conference on Harmonisation (ICH) guidance 415

13.2.2 US FDA process validation guidance 416

13.3 Process analytical technology (PAT) 417

13.3.1 At-line, online, inline or offline: what is the difference? 417

13.3.2 Enablers of PAT 419

13.4 The link between QbD and PAT 425

13.5 Chemometrics: the glue that holds QbD and PAT together 427

13.5.1 A new approach to batch process understanding: relative time modelling 428

13.5.2 Hierarchical modelling 434

13.5.3 Classification–classification hierarchies 434

13.5.4 Classification–prediction hierarchies 435

13.5.5 Prediction–prediction hierarchies 437

13.5.6 Continuous pharmaceutical manufacturing: the embodiment of QbD and PAT 438

13.6 An introduction to multivariate statistical process control (MSPC) 440

13.6.1 Aspects of data fusion 441

13.6.2 Multivariate statistical process control (MSPC) principles 443

13.6.3 Total process measurement system quality control (TPMSQC) 444

13.7 Model lifecycle management..445

13.7.1 The iterative model building cycle.. 446

13.7.2 A general procedure for model updating.. 448

13.7.3 Summary of model lifecycle management.. 449

13.8 Chapter summary..449

13.9 References..452