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
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
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>The first principal component</td>
<td>73</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Maximum variance directions</td>
<td>73</td>
</tr>
<tr>
<td>4.5.2</td>
<td>The first principal component as a least squares fit</td>
<td>74</td>
</tr>
<tr>
<td>4.6</td>
<td>Extension to higher-order principal components</td>
<td>75</td>
</tr>
<tr>
<td>4.7</td>
<td>Principal component models—scores and loadings</td>
<td>76</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Maximum number of principal components</td>
<td>76</td>
</tr>
<tr>
<td>4.7.2</td>
<td>PC model centre</td>
<td>77</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Introducing loadings—relations between X and PCs</td>
<td>77</td>
</tr>
<tr>
<td>4.7.4</td>
<td>Scores—coordinates in PC space</td>
<td>78</td>
</tr>
<tr>
<td>4.7.5</td>
<td>Object residuals</td>
<td>78</td>
</tr>
<tr>
<td>4.8</td>
<td>Objectives of PCA</td>
<td>79</td>
</tr>
<tr>
<td>4.9</td>
<td>Score plot–object relationships</td>
<td>80</td>
</tr>
<tr>
<td>4.9.1</td>
<td>Interpreting score plots</td>
<td>80</td>
</tr>
<tr>
<td>4.9.2</td>
<td>Choice of score plots</td>
<td>82</td>
</tr>
<tr>
<td>4.10</td>
<td>The loading plot–variable relationships</td>
<td>83</td>
</tr>
<tr>
<td>4.10.1</td>
<td>Correlation loadings</td>
<td>84</td>
</tr>
<tr>
<td>4.10.2</td>
<td>Comparison of scores and loading plots</td>
<td>86</td>
</tr>
<tr>
<td>4.10.3</td>
<td>The 1-dimensional loading plot</td>
<td>87</td>
</tr>
<tr>
<td>4.11</td>
<td>Example: city temperatures in europe</td>
<td>89</td>
</tr>
<tr>
<td>4.11.1</td>
<td>Introduction</td>
<td>89</td>
</tr>
<tr>
<td>4.11.2</td>
<td>Plotting data and deciding on the validation scheme</td>
<td>89</td>
</tr>
<tr>
<td>4.11.3</td>
<td>PCA results and interpretation</td>
<td>90</td>
</tr>
<tr>
<td>4.12</td>
<td>Principal component models</td>
<td>93</td>
</tr>
<tr>
<td>4.12.1</td>
<td>The PC model</td>
<td>93</td>
</tr>
<tr>
<td>4.12.2</td>
<td>Centring</td>
<td>93</td>
</tr>
<tr>
<td>4.12.3</td>
<td>Step by step calculation of PCs</td>
<td>94</td>
</tr>
<tr>
<td>4.12.4</td>
<td>A preliminary comment on the algorithm: NIPALS</td>
<td>94</td>
</tr>
<tr>
<td>4.12.5</td>
<td>Residuals—the E-matrix</td>
<td>95</td>
</tr>
<tr>
<td>4.12.6</td>
<td>Residual variance</td>
<td>95</td>
</tr>
<tr>
<td>4.12.7</td>
<td>Object residuals</td>
<td>96</td>
</tr>
<tr>
<td>4.12.8</td>
<td>The total squared object residual</td>
<td>96</td>
</tr>
<tr>
<td>4.12.9</td>
<td>Explained/residual variance plots</td>
<td>96</td>
</tr>
<tr>
<td>4.12.10</td>
<td>How many PCs to use?</td>
<td>97</td>
</tr>
<tr>
<td>4.12.11</td>
<td>A note on the number of PCs</td>
<td>98</td>
</tr>
<tr>
<td>4.12.12</td>
<td>A doubtful case—using external evidence</td>
<td>98</td>
</tr>
<tr>
<td>4.12.13</td>
<td>Variable residuals</td>
<td>99</td>
</tr>
</tbody>
</table>
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
Chapter 10. Supervised Classification Methods

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

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
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
Preface

The field of chemometrics is the application of multivariate data analysis (MVA) methodology to solve chemistry-based problems, and it has developed into a mature scientific discipline since its start in the early 80’s. Although there are numerous applications in daily use in many industries, the knowledge and use of methods are not so widespread. While the academic MVA research toolbox is well established and ever increasing, its industrial counterpart has slowly gained momentum, particularly over the last decade. This could be due to many reasons, such as the advancement of industrial data management solutions, the semi-continuous financial crises that drive industry to smarter manufacturing paradigms and intelligent data mining to reduce costs and waste. Whatever the reasons, it is clear that chemometrics and multivariate analysis have a big role to play in connecting the dots between research and the new industrial paradigm with advanced and more robust sensors, agile manufacturing processes and extensive product quality control. The necessity of proper procedures for collecting and analysing data has become even more important as everyone is into “Big Data” these days; however, a better term might be “Smart Data”. Welcome to the multivariate world!

This book aims at conveying a universal philosophy within the field of multivariate data analysis. The authors believe that all processes or systems are multivariate in nature until proven otherwise, and therefore must be analysed, modelled and understood as such. The topic of proper data quality is emphasised. There are two important aspects in any field that requires experimentation—the ability to understand the outputs of the experiment and the ability to put this newfound knowledge into use, for business gain or deeper scientific understanding. Whether the experiment involves measurements generated by scientific instruments, sensory data, manufacturing process data or psychometric variables, the two main questions to be answered of data analysis are: can the process/experiment be understood and how can this data be turned into useful information? This book will, first, arm the reader with the multivariate tools available to understand the data they generate and develop valid and robust models from such data and, second, show how these models can be applied in real-world situations. Chapter one, Introduction, elaborates on the above themes in full depth. In particular, the data analysis scope presented in this book a.o. is also a warning against too trigger-happy machine learning forays—the data analyst must be in charge.

The revision of this book includes an extended chapter on validation. Over the years, it has become clear that validation as a concept is not presented explicitly in most educational institutions. One reason might be that it does not belong in a specific science subject and therefore “falls between many stools”. The same can also be said about the principles of Design of Experiments (DoE), although its usage should be mandatory for anyone performing trials or observing processes. DoE is an important tool for understanding causality in a system and for confirming/rejecting test hypotheses. In combination with multivariate analysis, the symbiosis of DoE
and MVA gives the best of both worlds. It is also a fundamental basis for chemometrics that the owner of the data should also analyse them, with subject matter and domain specific background knowledge in mind.

The basic topics in the 5th edition have all been extended significantly, making the book thoroughly revised, updated and modernised, a.o. with a comprehensive chapter on basic statistics, a long-needed fundamental chapter on sampling which allows a new holistic view on both data quality and on the concept of “replication”. Finally, the book now contains a tour de force introduction to Process Analytical Technology (PAT) and Quality by Design (QbD) implementation. Although PAT has a strong connection to the pharmaceutical industry, the principles apply with equal force to many other application areas in technology and industry.

While the multivariate methods lend themselves to empirical analysis of data sampled from science, technology and nature, i.e. any system with multiple underlying structures, there is nothing that prevents the use of first principle models in combination with actual observations.

CAMO Software’s philosophy is not to describe all methods in the world into its Unscrambler® platform, but to provide methods that are versatile and suited for any kind of data, regardless of their size and properties. The authors and CAMO believe that the focus should be on graphical presentation of results and their interpretation rather than tables with p-values. This is related to the distinction between significance and relevance. With a high number of objects any test for significance between two groups or correlation between two variables will be statistically significant. Thus, a table of p-values does not necessarily show if a model is suitable for predicting selected properties such as product quality at the individual level.

When this is said and done, it is realised that summarising the important findings from a project or study is often efficiently done with bullet points or univariate statistics. Our message is that the multivariate methods provide the fastest insight into complex data to arrive at the correct conclusions and to avoid “searching for correlations”. The situation is that even after 40 years of multivariate methods and particularly multivariate calibration, it is not known to the majority of people that selectivity is not needed to predict quality of a product or classify/identify samples such as raw materials.

Being a data analyst is about practicing the methods and software on your own data.

We wish you all the best, and may your models be with you!

Prior to this edition of the book, previous editions included exercises with detailed description of how to operate The Unscrambler® software package. In this edition, however, the focus is on the analytical methods, their interpretation and finally, a topic that is not well covered to date, implementation of models including their proper use. For more details, see www.camo.com or videos on YouTube.

The 6th edition of this textbook/software education package represents a fundamentally updated, revised, extended and augmented new product. The present textbook is a result of an inspiring and pleasant joint effort of the two main authors in every aspect involved: didactic concept and design, revising trusted old chapter cores, de novo writing of ~50% new material, production of new illustrations and editorial subject-matter proofing. The final result owes a great debt of gratitude to two external reviewers with extensive didactic chemometric experience. We
express our most sincere gratitude for their comprehensive treatment of the first draft manuscript with well-reflected suggestions for an optimised learning “flow” as well as sharp insights concerning critical presentation details.

This book is now in its 6th edition (2017), being released on the basis of a total of 33,000 earlier copies; there is also a Russian language version (abbreviated contents). For the first author, it has been an immense privilege to serve the chemometric community, and indeed beyond, with this book since 1994 (1st Edition). The present 6th edition is—finally—one with which to be truly satisfied. A profound thank you goes to Brad for stepping in with gusto!

Kim H. Esbensen, Brad Swarbrick (authors—with a mission)
Frank Westad (contributing author, CAMO CSO)
Copenhagen, Sydney, Oslo, 1 August 2017