Contents

General notation 1

Abbreviations and acronyms 3

I Preface 7

1 Justification, objectives and contributions 9

1.1 An overview of the research project 9

1.2 Objectives of the thesis 10

1.2.1 Kernel-based methodologies for statistical process monitoring, improved fault diagnosis, translation of *out-of-control* signals to operator actions, and analysis of mixture designs of experiments 10

1.2.2 Rational approaches for selecting the optimal amount of information to be modelled for data exploration and understanding 11

1.2.3 Model transfer between manufacturing units or workstations 11

1.2.4 Real-time data processing 12

1.3 Contributions. 12

2 On latent variable-based multivariate data analysis 17

2.1 Introduction 17

2.2 Latent variable-based multivariate data analysis techniques 19

2.2.1 Principal Component Analysis (PCA) 19

2.2.2 Partial Least Squares regression (PLS) 20

2.2.3 Partial Least Squares Discriminant Analysis (PLSDA) 21

2.3 Some important additional notions: *cross-validation*, *jackknifing* and *permutation testing* 21

3 Materials and methods 23

3.1 Hardware 23

3.2 Software 23

3.3 Datasets and methods 23

II On kernel-based extensions of PCA, PLS and PLSDA 25

4 Preliminary considerations 27
4.1 Introduction 28
4.1.1 Kernel-based techniques: basic principles 29
4.1.2 Pseudo-samples and pseudo-sample projection 29

5 K-PLSDA for RGB image segmentation 33
5.1 Introduction 34
5.2 Methods 36
5.2.1 Colour analysis-based segmentation techniques 36
5.2.2 Texture analysis-based segmentation techniques 37
5.2.3 Graph-based segmentation techniques 38
5.2.4 Multivariate Image Analysis (MIA)-based segmentation techniques . . 38
5.3 Dataset 39
5.4 Comparative study 40
5.5 Results 41
5.6 Illustration case 47
5.7 Concluding remarks 51

6 K-PLSDA and pseudo-sample projection for batch run discrimination 53
6.1 Introduction 54
6.2 Methods 54
6.3 Datasets 55
6.4 Results and discussion 56
6.4.1 Simulated dataset 56
6.4.2 VWU/K-PLSDA (polymerisation process dataset) 61
6.4.3 BWU/K-PLSDA (polymerisation process dataset) 63
6.4.4 LFE/K-PLSDA (pharmaceutical batch drying process dataset) 67
6.5 Comparison between K-PLSDA and classical PLSDA models 70
6.6 Conclusions 71

7 K-PCA and pseudo-sample projection for batch process monitoring 73
7.1 Introduction 74
7.2 Adaptation of the pseudo-sample projection strategy to batch process monitoring 74
7.3 Datasets 77
7.4 Results and discussion 78
7.4.1 Simulated dataset - Variability increase detection case study 78
7.4.2 Simulated dataset - Variability decrease detection case study 89
7.4.3 Chemical process dataset 94
7.4.4 Pharmaceutical batch drying process dataset 100

7.5 Conclusions 101

8 K-PLS and pseudo-sample trajectories for mixture data analysis 103

8.1 Introduction 104

8.2 Methods 105

8.2.1 Scheffé and Cox models 105

8.2.2 Pseudo-sample trajectories for mixture data 106

8.2.3 Pseudo-sample-based response surfaces 108

8.3 Datasets 109

8.3.1 Simulated data 109

8.3.2 Tablet data 109

8.3.3 Bubbles data 110

8.3.4 Colorant data 110

8.3.5 Photographic paper data 110

8.4 Results 110

8.4.1 Simulated data 110

8.4.2 Tablet data 112

8.4.3 Bubbles data 114

8.4.4 Colorant data 117

8.4.5 Photographic paper data 120

8.5 Conclusions 120

III On the selection of the number of factors in PCA by permutation testing 125

9 A novel permutation test-based approach for PCA component selection 127

9.1 Introduction 128

9.1.1 Strategies for principal component selection 128

9.2 Methodology 129

9.3 Theoretical and practical aspects of the algorithm 132

9.3.1 Permutations 133

9.3.2 Deflation 133

9.3.3 Projection 134

9.3.4 The rationale behind F_a 138

9.4 Performance of the algorithm 139

9.4.1 Synthetic datasets 139

9.4.2 Real case studies 143

9.5 Conclusions 148

IV On modelling common and distinctive sources of variability in multi-set data analysis 149

10 Some considerations on two-block common and distinctive component analysis 151

10.1 Introduction 152
10.2 Methods 154
10.2.1 Simultaneous Component Analysis (SCA) 154
10.2.2 DIStinctive and COmmon Simultaneous Component Analysis (DISCO-SCA) 155
10.2.3 Adapted Generalised Singular Value Decomposition (Adapted GSVD) 155
10.2.4 ECO-POWER 156
10.2.5 Canonical Correlation Analysis (CCA) 156
10.2.6 2-block Orthogonal Projections to Latent Structures (O2PLS) 157
10.3 Datasets 158
10.3.1 Gene expression data 158
10.3.2 Simulated pseudo-spectral data 158
10.3.3 Industrial batch process data 159
10.4 Results and discussion 159
10.4.1 Is the variation accounted for reliable? 159
10.4.2 Determining the number of common and distinctive components: a novel strategy 161
10.4.3 Modelling common and distinctive components in regression scenarios 169
10.5 Conclusions 174

11 Calibration transfer between near-infrared spectrometers 175

11.1 Introduction 176
11.2 Methods 177
11.2.1 Piecewise Direct Standardisation (PDS) 177
11.2.2 Maximum Likelihood Principal Component Analysis (MLPCA) . . . 177
11.2.3 Trimmed Scores Regression (TSR) 178
11.2.4 JYPLS-based approaches 179
11.3 Modelling procedure 180
11.4 Datasets 182
11.5 Results 183
11.5.1 Gasoline dataset 183
11.5.2 Corn dataset 187
11.6 Discussion 190
11.7 Conclusions 191

V On the on-the fly processing and modelling of continuous high-dimensional data streams 193

12 The On-The-Fly Processing tool 195

12.1 Introduction 196
12.1.1 Data compression strategies 196
12.1.2 Subspace compression 198
12.1.3 PCA as a multivariate series expansion of the underlying data generation mechanism 198
12.1.4 Algorithms for PCA decomposition 199
12.2 System overview 200
12.2.1 Input 202
12.2.2 Fit to already established model subspace 203
12.2.3 Bilinear model expansion 204
12.2.4 Model updating 204
12.3 Datasets 205
12.4 Results and discussion 206
12.4.1 High-speed multi-channel monitoring of the Belousov-Zhabotinsky reaction 206
12.4.2 Detailed remote characterisation of orange samples 208
12.4.3 Environmental surveillance by airborne hyperspectral imaging 211
12.4.4 Analysis of an industrial manufacturing process 213
12.5 Comparison with classical PCA 215
12.6 Discussion 216
12.7 Conclusions and perspectives 217

VI Epilogue 219

13 Conclusions and perspectives 221

13.1 Accomplishment of the objectives 221
13.1.1 Objective I - Exploring the potential of kernel-based methodologies for statistical process monitoring, improved fault diagnosis, translation of *out-of-control signals* to operator actions, and analysis of mixture designs of experiments 221
13.1.2 Objective II - Proposing rational approaches for selecting the optimal amount of information to be modelled for data exploration and understanding 222
13.1.3 Objective III - Enhancing model transfer between manufacturing units or workstations. 222
13.1.4 Objective IV - Implementing new computational strategies for real-time data processing 223
13.2 Future research lines 224

14 Appendices **227**

14.1 Annex to Part II 227

14.1.1 Relationship between the Euclidean distance matrix, $\mathbf{D}$, and the inner product matrix, $\mathbf{XX}^{\mathrm{T}}$ 227

14.1.2 Practical meaning of the pseudo-samples in the feature space 228

14.1.3 Relationship between Scheffé and Cox model coefficients 229

14.2 Annex to Part III 231

14.2.1 Horn's parallel analysis 231

14.2.2 Dray's method 231

14.3 Annex to Part IV 233

14.3.1 The JYPLS algorithm 233

14.3.2 Principal Component Regression (PCR) 235

14.4 Annex to Part V 235

14.4.1 Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) 235

Bibliography **237**