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

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

5 K-PLSDA for RGB image segmentation  
5.1 Introduction  
5.2 Methods  
5.2.1 Colour analysis-based segmentation techniques  
5.2.2 Texture analysis-based segmentation techniques  
5.2.3 Graph-based segmentation techniques  
5.2.4 Multivariate Image Analysis (MIA)-based segmentation techniques  
5.3 Dataset  
5.4 Comparative study  
5.5 Results  
5.6 Illustration case  
5.7 Concluding remarks

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

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

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

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

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

The On-The-Fly Processing tool

1.1 Introduction

1.1.1 Data compression strategies

1.1.2 Subspace compression

1.1.3 PCA as a multivariate series expansion of the underlying data generation mechanism

1.1.4 Algorithms for PCA decomposition

1.2 System overview

1.2.1 Input

1.2.2 Fit to already established model subspace

1.2.3 Bilinear model expansion

1.2.4 Model updating

1.3 Datasets

1.4 Results and discussion

1.4.1 High-speed multi-channel monitoring of the Belousov-Zhabotinsky reaction

1.4.2 Detailed remote characterisation of orange samples

1.4.3 Environmental surveillance by airborne hyperspectral imaging

1.4.4 Analysis of an industrial manufacturing process

1.5 Comparison with classical PCA

1.6 Discussion

1.7 Conclusions and perspectives

Conclusion and perspectives

1.1 Accomplishment of the objectives

1.1.1 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

1.1.2 Objective II - Proposing rational approaches for selecting the optimal amount of information to be modelled for data exploration and understanding

1.1.3 Objective III - Enhancing model transfer between manufacturing units or workstations

1.1.4 Objective IV - Implementing new computational strategies for real-time data processing

1.2 Future research lines
## 14 Appendices

14.1 Annex to Part II ............................................. 227
   14.1.1 Relationship between the Euclidean distance matrix, $D$, and the
   inner product matrix, $XX^T$ .................................. 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