

---

# Contents

<b>1</b>	<b>Introduction</b> .....	<b>1</b>
----------	---------------------------	----------

---

## Part I Preliminaries

---

<b>2</b>	<b>Data</b> .....	<b>7</b>
<b>3</b>	<b>Preprocessing</b> .....	<b>13</b>
3.1	Dealing with Noise .....	13
3.2	Baseline Removal .....	18
3.3	Aligning Peaks – Warping .....	20
3.3.1	Parametric Time Warping .....	22
3.3.2	Dynamic Time Warping .....	26
3.3.3	Practicalities .....	31
3.4	Peak Picking .....	31
3.5	Scaling .....	33
3.6	Missing Data .....	38
3.7	Conclusion .....	39

---

## Part II Exploratory Analysis

---

<b>4</b>	<b>Principal Component Analysis</b> .....	<b>43</b>
4.1	The Machinery .....	44
4.2	Doing It Yourself .....	46
4.3	Choosing the Number of PCs .....	48
4.3.1	Statistical Tests .....	49
4.4	Projections .....	51
4.5	R Functions for PCA .....	53
4.6	Related Methods .....	57
4.6.1	Multidimensional Scaling .....	57

4.6.2	Independent Component Analysis and Projection Pursuit .....	60
4.6.3	Factor Analysis .....	63
4.6.4	Discussion .....	65
<b>5</b>	<b>Self-Organizing Maps</b> .....	<b>67</b>
5.1	Training SOMs .....	68
5.2	Visualization .....	71
5.3	Application .....	73
5.4	R Packages for SOMs .....	76
5.5	Discussion .....	77
<b>6</b>	<b>Clustering</b> .....	<b>79</b>
6.1	Hierarchical Clustering .....	80
6.2	Partitional Clustering .....	85
6.2.1	K-Means .....	85
6.2.2	K-Medoids .....	87
6.3	Probabilistic Clustering .....	90
6.4	Comparing Clusterings .....	95
6.5	Discussion .....	97

---

## Part III Modelling

---

<b>7</b>	<b>Classification</b> .....	<b>103</b>
7.1	Discriminant Analysis .....	104
7.1.1	Linear Discriminant Analysis .....	105
7.1.2	Crossvalidation .....	109
7.1.3	Fisher LDA .....	111
7.1.4	Quadratic Discriminant Analysis .....	114
7.1.5	Model-Based Discriminant Analysis .....	116
7.1.6	Regularized Forms of Discriminant Analysis .....	118
7.2	Nearest-Neighbour Approaches .....	122
7.3	Tree-Based Approaches .....	126
7.3.1	Recursive Partitioning and Regression Trees .....	126
7.3.2	Discussion .....	135
7.4	More Complicated Techniques .....	135
7.4.1	Support Vector Machines .....	136
7.4.2	Artificial Neural Networks .....	141
<b>8</b>	<b>Multivariate Regression</b> .....	<b>145</b>
8.1	Multiple Regression .....	145
8.1.1	Limits of Multiple Regression .....	147
8.2	PCR .....	149
8.2.1	The Algorithm .....	149

8.2.2	Selecting the Optimal Number of Components . . . . .	152
8.3	Partial Least Squares (PLS) Regression . . . . .	155
8.3.1	The Algorithm(s) . . . . .	156
8.3.2	Interpretation . . . . .	160
8.4	Ridge Regression . . . . .	163
8.5	Continuum Methods . . . . .	165
8.6	Some Non-Linear Regression Techniques . . . . .	165
8.6.1	SVMs for Regression . . . . .	165
8.6.2	ANNs for Regression . . . . .	168
8.7	Classification as a Regression Problem . . . . .	170
8.7.1	Regression for LDA . . . . .	170
8.7.2	Discussion . . . . .	172

---

## Part IV Model Inspection

---

9	Validation . . . . .	175
9.1	Representativity and Independence . . . . .	176
9.2	Error Measures . . . . .	178
9.3	Model Selection . . . . .	179
9.4	Crossvalidation Revisited . . . . .	181
9.4.1	LOO Crossvalidation . . . . .	181
9.4.2	Leave-Multiple-Out Crossvalidation . . . . .	183
9.4.3	Double Crossvalidation . . . . .	183
9.5	The Jackknife . . . . .	184
9.6	The Bootstrap . . . . .	186
9.6.1	Error Estimation with the Bootstrap . . . . .	187
9.6.2	Confidence Intervals for Regression Coefficients . . . . .	190
9.6.3	Other R Packages for Bootstrapping . . . . .	195
9.7	Integrated Modelling and Validation . . . . .	195
9.7.1	Bagging . . . . .	196
9.7.2	Random Forests . . . . .	197
9.7.3	Boosting . . . . .	202
10	Variable Selection . . . . .	205
10.1	Tests for Coefficient Significance . . . . .	206
10.1.1	Confidence Intervals for Individual Coefficients . . . . .	207
10.1.2	Tests Based on Overall Error Contributions . . . . .	210
10.2	Explicit Coefficient Penalization . . . . .	213
10.3	Global Optimization Methods . . . . .	217
10.3.1	Simulated Annealing . . . . .	218
10.3.2	Genetic Algorithms . . . . .	225
10.3.3	Discussion . . . . .	232

---

**Part V Applications**

---

<b>11 Chemometric Applications</b> . . . . .	235
11.1 Outlier Detection with Robust PCA . . . . .	235
11.1.1 Robust PCA . . . . .	236
11.1.2 Discussion . . . . .	240
11.2 Orthogonal Signal Correction and OPLS . . . . .	240
11.3 Discrimination with Fat Data Matrices . . . . .	243
11.3.1 PCDA . . . . .	244
11.3.2 PLSDA . . . . .	248
11.4 Calibration Transfer . . . . .	251
11.5 Multivariate Curve Resolution . . . . .	255
11.5.1 Theory . . . . .	256
11.5.2 Finding Suitable Initial Estimates . . . . .	257
11.5.3 Applying MCR . . . . .	261
11.5.4 Constraints . . . . .	263
11.5.5 Combining Data Sets . . . . .	265

---

**Part VI Appendices**

---

<b>R Packages Used in this Book</b> . . . . .	271
<b>References</b> . . . . .	273
<b>Index</b> . . . . .	283