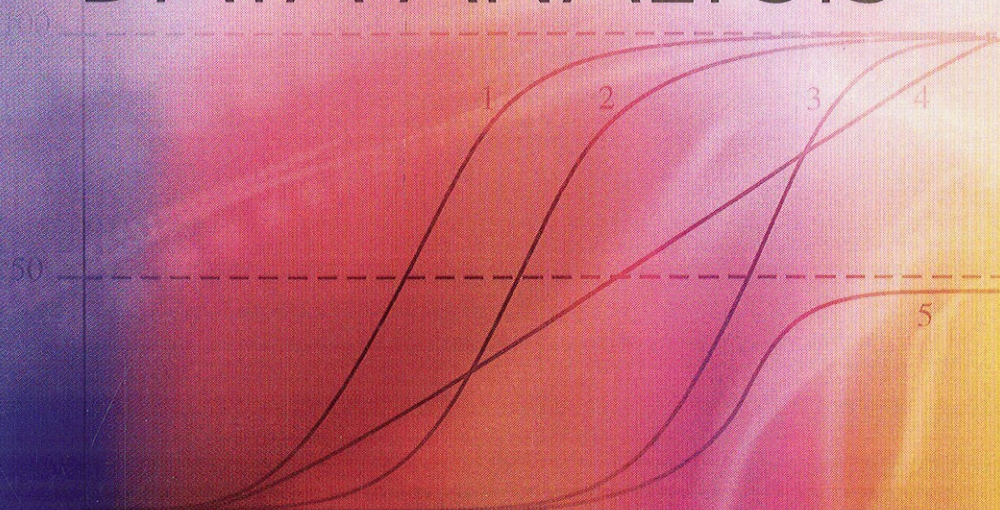
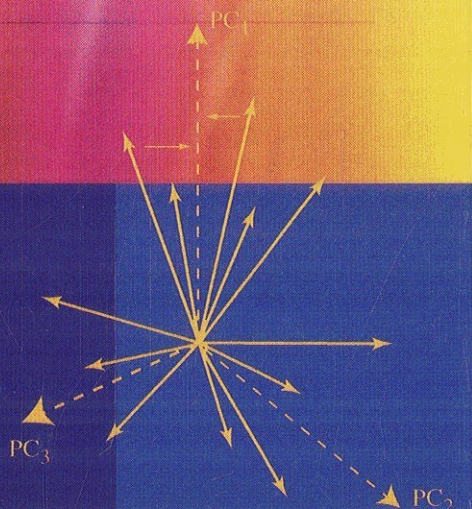


A PRACTICAL GUIDE TO SCIENTIFIC DATA ANALYSIS



DAVID LIVINGSTONE

 WILEY



Contents

Preface	xi
Abbreviations	xiii
1 Introduction: Data and Its Properties, Analytical Methods and Jargon	1
1.1 Introduction	2
1.2 Types of Data	3
1.3 Sources of Data	5
1.3.1 Dependent Data	5
1.3.2 Independent Data	6
1.4 The Nature of Data	7
1.4.1 Types of Data and Scales of Measurement	8
1.4.2 <i>Data Distribution</i>	10
1.4.3 Deviations in Distribution	15
1.5 Analytical Methods	19
1.6 Summary	23
References	23
2 Experimental Design – Experiment and Set Selection	25
2.1 What is Experimental Design?	25
2.2 Experimental Design Techniques	27
2.2.1 Single-factor Design Methods	31
2.2.2 Factorial Design (Multiple-factor Design)	33
2.2.3 D-optimal Design	38
2.3 Strategies for Compound Selection	40
2.4 High Throughput Experiments	51
2.5 Summary	53
References	54

3	Data Pre-treatment and Variable Selection	57
3.1	Introduction	57
3.2	Data Distribution	58
3.3	Scaling	60
3.4	Correlations	62
3.5	Data Reduction	63
3.6	Variable Selection	67
3.7	Summary	72
	References	73
4	Data Display	75
4.1	Introduction	75
4.2	Linear Methods	77
4.3	Nonlinear Methods	94
4.3.1	Nonlinear Mapping	94
4.3.2	Self-organizing Map	105
4.4	Faces, Flowerplots and Friends	110
4.5	Summary	113
	References	116
5	Unsupervised Learning	119
5.1	Introduction	119
5.2	Nearest-neighbour Methods	120
5.3	Factor Analysis	125
5.4	Cluster Analysis	135
5.5	Cluster Significance Analysis	140
5.6	Summary	143
	References	144
6	Regression Analysis	145
6.1	Introduction	145
6.2	Simple Linear Regression	146
6.3	Multiple Linear Regression	154
6.3.1	Creating Multiple Regression Models	159
6.3.1.1	Forward Inclusion	159
6.3.1.2	Backward Elimination	161
6.3.1.3	Stepwise Regression	163
6.3.1.4	All Subsets	164
6.3.1.5	Model Selection by Genetic Algorithm	165
6.3.2	Nonlinear Regression Models	167
6.3.3	Regression with Indicator Variables	169

CONTENTS

6.4	Multiple Regression: Robustness, Chance Effects, the Comparison of Models and Selection Bias	174
6.4.1	Robustness (Cross-validation)	174
6.4.2	Chance Effects	177
6.4.3	Comparison of Regression Models	178
6.4.4	Selection Bias	180
6.5	Summary	183
	References	184
7	Supervised Learning	187
7.1	Introduction	187
7.2	Discriminant Techniques	188
7.2.1	Discriminant Analysis	188
7.2.2	SIMCA	195
7.2.3	Confusion Matrices	198
7.2.4	Conditions and Cautions for Discriminant Analysis	201
7.3	Regression on Principal Components and PLS	202
7.3.1	Regression on Principal Components	203
7.3.2	Partial Least Squares	206
7.3.3	Continuum Regression	211
7.4	Feature Selection	214
7.5	Summary	216
	References	217
8	Multivariate Dependent Data	219
8.1	Introduction	219
8.2	Principal Components and Factor Analysis	221
8.3	Cluster Analysis	230
8.4	Spectral Map Analysis	233
8.5	Models with Multivariate Dependent and Independent Data	238
8.6	Summary	246
	References	247
9	Artificial Intelligence and Friends	249
9.1	Introduction	250
9.2	Expert Systems	251
9.2.1	LogP Prediction	252
9.2.2	Toxicity Prediction	261
9.2.3	Reaction and Structure Prediction	268

9.3	Neural Networks	273
9.3.1	Data Display Using ANN	277
9.3.2	Data Analysis Using ANN	280
9.3.3	Building ANN Models	287
9.3.4	Interrogating ANN Models	292
9.4	Miscellaneous AI Techniques	295
9.5	Genetic Methods	301
9.6	Consensus Models	303
9.7	Summary	304
	References	305
10	Molecular Design	309
10.1	The Need for Molecular Design	309
10.2	What is QSAR/QSPR?	310
10.3	Why Look for Quantitative Relationships?	321
10.4	Modelling Chemistry	323
10.5	Molecular Fields and Surfaces	325
10.6	Mixtures	327
10.7	Summary	329
	References	330
	Index	333