Contents

I Biostatistics **1**

1 Discriminant Analysis Based on Continuous and Discrete Variables **3**

Avner Bar-Hen and Jean-Jacques Daudin

1.1 Introduction 3

1.2 Generalisation of the Mahalanobis Distance 4

1.2.1 Introduction 4

1.2.2 Kullback–Leibler Divergence 5

1.2.3 Asymptotic Distribution of Matusita Distance 10

1.2.4 Simulations 12

1.3 Methods and Stopping Rules for Selecting Variables 13

1.4 Reject Option 15

1.4.1 Distributional Result 15

1.4.2 Derivation of the Preliminary Test 18

1.5 Example 22

1.5.1 Location Model 22

1.5.2 Comparison with the Linear Discriminant Analysis . . 24

1.5.3 Conclusion 24

Bibliography 25

2 Longitudinal Data Analysis with Linear Regression 29

Jörg Breitung, Rémy Slama and Axel Werwatz

2.1 Introduction 29
2.2 Theoretical Aspects 32
2.2.1 The Fixed-effect Model 32
2.2.2 The Random Effects Model 36
2.3 Computing Fixed and Random-effect Models 37
2.3.1 Data Preparation 37
2.3.2 Fixed and Random-effect Linear Regression 38
2.3.3 Options for `panfix` 38
2.3.4 Options for `panrand` 39
2.4 Application 40
2.4.1 Results 41
Bibliography 43

3 A Kernel Method Used for the Analysis of Replicated Micro-array Experiments 45

Ali Gannoun, Beno Liquetît, Jérôme Saracco and Wolfgang Urfer

3.1 Introduction 45
3.2 Statistical Model and Some Existing Methods 46
3.2.1 The Basic Model 47
3.2.2 The T-test 47
3.2.3 The Mixture Model Approach 48
3.3 A Fully Nonparametric Approach 49
3.3.1 Kernel Estimation of f_0 and f 50
3.3.2 The Reflection Approach in Kernel Estimation 50
3.3.3 Implementation of the Nonparametric Method 51
3.4 Data Analysis 52
3.4.1 Results Obtained with the Normal Mixture Model . . 53

3.4.2 Results Obtained with the Nonparametric Approach . 53
3.4.3 A Simulation Study 56
3.5 Discussion and Concluding Remarks 58
Bibliography 59

4 Kernel Estimates of Hazard Functions for Biomedical Data Sets 63

Ivana Horová and Jiří Zelinka

4.1 Introduction 63
4.2 Kernel Estimate of the Hazard Function and Its Derivatives . 64
4.3 Choosing the Shape of the Kernel 68
4.4 Choosing the Bandwidth 69
4.5 Description of the Procedure 74
4.6 Application 75
Bibliography 83

5 Partially Linear Models 87

Wolfgang Härdle and Hua Liang

5.1 Introduction 87
5.2 Estimation and Nonparametric Fits 89
5.2.1 Kernel Regression 89
5.2.2 Local Polynomial 90
5.2.3 Piecewise Polynomial 93
5.2.4 Least Square Spline 96
5.3 Heteroscedastic Cases 97
5.3.1 Variance Is a Function of Exogenous Variables 98
5.3.2 Variance Is an Unknown Function of T 99
5.3.3 Variance Is a Function of the Mean 99
5.4 Real Data Examples 100
Bibliography 102

6 Analysis of Contingency Tables **105**

Masahiro Kuroda

6.1 Introduction 105
6.2 Log-linear Models 105
6.2.1 Log-linear Models for Two-way Contingency Tables . 106
6.2.2 Log-linear Models for Three-way Contingency Tables . 107
6.2.3 Generalized Linear Models 109
6.2.4 Fitting to Log-linear Models 111
6.3 Inference for Log-linear Models Using XploRe 113
6.3.1 Estimation of the Parameter Vector λ 113
6.3.2 Computing Statistics for the Log-linear Models 113
6.3.3 Model Comparison and Selection 114
6.4 Numerical Analysis of Contingency Tables 115
6.4.1 Testing Independence 115
6.4.2 Model Comparison 119
Bibliography 124

7 Identifying Coexpressed Genes **125**

Qihua Wang

7.1 Introduction 125
7.2 Methodology and Implementation 127
7.2.1 Weighting Adjustment 128
7.2.2 Clustering 132
7.3 Concluding Remarks 142
Bibliography 144

8 Bootstrap Methods for Testing Interactions in GAMs 147

Javier Roca-Pardiñas, Carmen Cadarso-Suárez and Wenceslao González-Manteiga

8.1 Introduction . . . 147
8.2 Logistic GAM with Interactions . . . 149
8.2.1 Estimation: the Local Scoring Algorithm . . . 150
8.3 Bootstrap-based Testing for Interactions . . . 152
8.3.1 Likelihood Ratio-based Test . . . 153
8.3.2 Direct Test . . . 153
8.3.3 Bootstrap Approximation . . . 153
8.4 Simulation Study . . . 154
8.5 Application to Real Data Sets . . . 156
8.5.1 Neural Basis of Decision Making . . . 156
8.5.2 Risk of Post-operative Infection . . . 159
8.6 Discussion . . . 162
8.7 Appendix . . . 163
Bibliography . . . 165

9 Survival Trees 167

Carmela Cappelli and Heping Zhang

9.1 Introduction . . . 167
9.2 Methodology . . . 170
9.2.1 Splitting Criteria . . . 170
9.2.2 Pruning . . . 173
9.3 The Quantlet stree . . . 174
9.3.1 Syntax . . . 174
9.3.2 Example . . . 175
Bibliography . . . 179

10 A Semiparametric Reference Curves Estimation **181**

Saracco Jérôme, Gannoun Ali, Guinot Christiane and Liquet Benoît

10.1 Introduction 181
10.2 Kernel Estimation of Reference Curves 184
10.3 A Semiparametric Approach Via Sliced Inverse Regression . . 187
10.3.1 Dimension Reduction Context 187
10.3.2 Estimation Procedure 191
10.3.3 Asymptotic Property 192
10.3.4 A Simulated Example 193
10.4 Case Study on Biophysical Properties of the Skin 195
10.4.1 Overview of the Variables 196
10.4.2 Methodological Procedure 197
10.4.3 Results and Interpretation 198
10.5 Conclusion 200
Bibliography 201

11 Survival Analysis **207**

Makoto Tomita

11.1 Introduction 207
11.2 Data Sets 208
11.3 Data on the Period up to Sympton Recurrence 208
11.3.1 Kaplan-Meier Estimate 208
11.3.2 log-rank Test 210
11.4 Data for Aseptic Necrosis 211
11.4.1 Kaplan-Meier Estimate 214
11.4.2 log-rank Test 214
11.4.3 Cox's Regression 215
Bibliography 217

II Related Sciences 219

12 Ozone Pollution Forecasting 221

Hervé Cardot, Christophe Crambes and Pascal Sarda

12.1 Introduction 221
12.2 A Brief Analysis of the Data 222
12.2.1 Description of the Data 222
12.2.2 Principal Component Analysis 224
12.2.3 Functional Principal Component Analysis 225
12.3 Functional Linear Model 226
12.3.1 Spline Estimation of α 227
12.3.2 Selection of the Parameters 228
12.3.3 Multiple Functional Linear Model 229
12.4 Functional Linear Regression for Conditional Quantiles Estimation 231
12.4.1 Spline Estimator of Ψ_α 232
12.4.2 Multiple Conditional Quantiles 234
12.5 Application to Ozone Prediction 235
12.5.1 Prediction of the Conditional Mean 236
12.5.2 Prediction of the Conditional Median 236
12.5.3 Analysis of the Results 238
Bibliography 242

13 Nonparametric Functional Chemometric Analysis 245

Frédéric Ferraty, Aldo Goia and Philippe Vieu

13.1 Introduction 245
13.2 General Considerations 246
13.2.1 Introduction to Spectrometric Data 246
13.2.2 Introduction to Nonparametric Statistics for Curves . 248

13.2.3 Notion of Proximity Between Curves 249
13.2.4 XploRe Quantlets for Proximity Between Curves . . . 251
13.3 Functional Nonparametric Regression 252
13.3.1 The Statistical Problem 252
13.3.2 The Nonparametric Functional Estimate 253
13.3.3 Prediction of Fat Percentage from Continuous Spectrum 254
13.3.4 The XploRe Quantlet 255
13.3.5 Comments on Bandwidth Choice 256
13.4 Nonparametric Curves Discrimination 257
13.4.1 The Statistical Problem 257
13.4.2 A Nonparametric Curves Discrimination Method . . . 258
13.4.3 Discrimination of Spectrometric Curves 260
13.4.4 The XploRe Quantlet 261
13.5 Concluding Comments 262
Bibliography 263

14 Variable Selection in Principal Component Analysis 265

Yuichi Mori, Masaya Iizuka, Tomoyuki Tarumi and Yutaka Tanaka

14.1 Introduction 265
14.2 Variable Selection in PCA 267
14.3 Modified PCA 268
14.4 Selection Procedures 269
14.5 Quantlet 272
14.6 Examples 273
14.6.1 Artificial Data 273
14.6.2 Application Data 279
Bibliography 282

15 Spatial Statistics **285**

Pavel Čížek, Wolfgang Härdle and Jürgen Symanzik

15.1 Introduction 285
15.2 Analysis of Geostatistical Data 287
15.2.1 Trend Surfaces 288
15.2.2 Kriging 290
15.2.3 Correlogram and Variogram 292
15.3 Spatial Point Process Analysis 297
15.4 Discussion 303
15.5 Acknowledgements 303
Bibliography 303

16 Functional Data Analysis **305**

Michal Benko

16.1 Introduction 305
16.2 Functional Basis Expansion 307
16.2.1 Fourier Basis 308
16.2.2 Polynomial Basis 309
16.2.3 B-Spline Basis 309
16.2.4 Data Set as Basis 309
16.3 Approximation and Coefficient Estimation 310
16.3.1 Software Implementation 312
16.3.2 Temperature Example 313
16.4 Functional Principal Components 314
16.4.1 Implementation 317
16.4.2 Data Set as Basis 319
16.5 Smoothed Principal Components Analysis 321
16.5.1 Implementation Using Basis Expansion 323
16.5.2 Temperature Example 323

Bibliography 326

17 Analysis of Failure Time with Microearthquakes Applications 329

Graciela Estévez-Pérez and Alejandro Quintela del Rio

17.1 Introduction 329
17.2 Kernel Estimation of Hazard Function 330
17.3 An Application to Real Data 336
17.3.1 The Occurrence Process of Earthquakes 336
17.3.2 Galicia Earthquakes Data 337
17.4 Conclusions 342
Bibliography 343

18 Landcover Prediction 347

Frédéric Ferraty, Martin Paegelow and Pascal Sarda

18.1 Introduction 347
18.2 Presentation of the Data 348
18.2.1 The Area: the Garrotxes 348
18.2.2 The Data Set 348
18.3 The Multilogit Regression Model 349
18.4 Penalized Log-likelihood Estimation 351
18.5 Polychotomous Regression in Action 352
18.6 Results and Interpretation 353
Bibliography 356

19 The Application of Fuzzy Clustering to Satellite Images Data 357

Hizir Sofyan, Muzailin Affan and Khaled Bawahidi

19.1 Introduction 357
19.2 Remote Sensing 358
19.3 Fuzzy C-means Method 359

19.3.1 Data and Methods 361
19.4 Results and Discussions 362
Bibliography 366

Index **367**