

Statistical Methods for Biostatistics and Related Fields

Wolfgang Härdle · Yuichi Mori
Philippe Vieu

Statistical Methods for Biostatistics and Related Fields

With 91 Figures and 57 Tables

 Springer

Professor Wolfgang Härdle
CASE - Center for Applied Statistics and Economics
Institute for Statistics and Econometrics
School of Business and Economics
Humboldt Universität zu Berlin
Spandauer Str. 1
10178 Berlin
haerdle@wiwi.hu-berlin.de

Professor Yuichi Mori
Department of Socio-information
Okayama University of Sciences
1-1 Ridai-cho
700-0005 Okayama
mori@soci.ous.ac.jp

Professor Philippe Vieu
Laboratoire de Statistique et Probabilités
Université Paul Sabatier
118 route de Narbonne
31062 Toulouse Cedex
vieu@cict.fr

Library of Congress Control Number: 2006934207

ISBN-13 978-3-540-32690-8 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Production: LE-TeX Jelonek, Schmidt & Vöckler GbR, Leipzig

Cover-design: Erich Kirchner, Heidelberg

SPIN 11681021

154/3100YL - 5 4 3 2 1 0

Printed on acid-free paper

Preface

Biostatistics is one of the scientific fields for which the developments during the last decades of the 20th century have been the most important. Biostatistics is a pluri-disciplinary area combining statistics and biology, but also agronomics, medicine or health sciences. It needs a good knowledge of the mathematical background inherent in statistical methodology, in order to understand the various fields of applications. The idea of this book is to present a variety of research papers on the state of art in modern biostatistics.

Biostatistics is interacting with many scientific fields. To highlight this wide *diversity*, we deliberately put these interactions at the center of our project. Our book is therefore divided into two parts. Part I is presenting several statistical models and methods for different biologic applications, while Part II will be concerned with problems and statistical methods coming from other related scientific fields.

This book intends to provide a basis for many people interested in biostatistics and related sciences. Students, teachers and academic researchers will find an overview on modelling and statistical analysis of biological data. Also, the book is meant for practitioners involved in research organisations (pharmacologic industry, medicine, food industry,..) for which statistics is an indispensable tool.

Biology is a science which has always been in permanent interaction with many other fields such as medicine, physics, environmetrics, chemistry, mathematics, probability, statistics On the other hand, statistics is interacting with many other fields of mathematics as with almost all other scientific disciplines, including biology. For all these reasons, biostatistics is strongly dependent on other scientific fields, and in order to provide a wide angle overview we present here a rich *diversity* of applied problems.

Each contribution of this book presents one (or more) real problem. The variation ranges from biological problems (see Chapter 1 and 10), medical contributions (see Chapters 2, 4, 5, 8, 9 or 11) and genomics contributions (see Chapters 3 and 7), to applications coming from other scientific areas, such as

environmetrics (see Chapters 12), chemometrics (see Chapter 13), geophysics (see Chapters 17 and 18) or image analysis (see Chapter 18). Because all these disciplines are continuously taking benefits one from each other, this choice highlights as well how each biostatistical method and modelling is helpful in other areas and vice versa.

A good illustration of such a duality is provided by hazard analysis, which is originally a medical survival problem (see Chapters 4, 9 or 11) but which leads to substantial interest in many other fields (see e.g. the microearthquakes analysis presented in Chapter 17). Another example is furnished by spatial statistics (see Chapters 15 or 18) or food industry problems (see Chapter 13), which are apparently far from medical purposes but whose developments have obvious (and strong) consequences in medical image analysis and in biochemical studies.

Due to the variety of applied biostatistical problems, the scope of methods is also very large. We address therefore the *diversity* of these statistical approaches by presenting recent developments in descriptive statistics (see Chapters 7, 9, 14 and 19), parametric modelling (see Chapters 1, 2, 6 and 18) nonparametric estimation (see Chapters 3, 4, 11, 15 and 17) and semi-parametrics (see Chapters 5, 8 and 10). An important place is devoted to methods for analyzing functional data (see Chapters 12, 13, 16), which is currently an active field of modern statistics.

An important feature of biostatistics is to have to deal with rather large statistical sample sizes. This is particular true for genomics applications (see Chapters 3 and 7) and for functional data modelling (see Chapters 12, 13 and 16). The computational issues linked with the methodologies presented in this book are carried out thanks to the capacities of the XploRe environment. Most of the methodological contributions are accompanied with automatic and/or interactive XploRe quantlets.

We would like to express our gratitude to all the contributors. We are confident that the scope of papers will insure a large impact of this book on future research lines and/or on applications in biostatistics and related fields. We would also like to express our sincere gratitude to all the researchers that we had the opportunity to meet in the past years. It would be tedious (and hardly exhaustive) to name all of them expressly here but specific thanks have to be addressed to our respective teams, will special mention to Anton Andriyashin in Berlin and to the participants of the STAPH working group in Toulouse.

July 2006
Berlin, Okoyama, Toulouse

*Wolfgang Härdle, Yuichi Mori
and Philippe Vieu*

Contents

I	Biostatistics	1
1	Discriminant Analysis Based on Continuous and Discrete Variables	3
	<i>Avner Bar-Hen and Jean-Jacques Daudin</i>	
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 <code>panfix</code>	38
2.3.4	Options for <code>panrand</code>	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 Liqueât, 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
	<i>Ivana Horová and Jiří Zelinka</i>	
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
	<i>Wolfgang Härdle and Hua Liang</i>	
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
	<i>Yuichi Mori, Masaya Iizuka, Tomoyuki Tarumi and Yutaka Tanaka</i>	
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
<i>Graciela Estévez-Pérez and Alejandro Quintela del Rio</i>	
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
<i>Frédéric Ferraty, Martin Paegelow and Pascal Sarda</i>	
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
<i>Hizir Sofyan, Muzailin Affan and Khaled Bawahidi</i>	
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

Part I

Biostatistics

1 Discriminant Analysis Based on Continuous and Discrete Variables

Avner Bar-Hen and Jean-Jacques Daudin

1.1 Introduction

In discrimination, as in many multivariate techniques, computation of a distance between two populations is often useful. For example in taxonomy, one can be interested not only in discriminating between two populations but in having an idea of how far apart the populations are. Mahalanobis' Δ^2 has become the standard measure of distance when the observations are quantitative and Hotelling derived its distribution for normal populations. The aim of this chapter is to adapt these results to the case where the observed characteristics are a mixture of quantitative and qualitative variables.

A problem frequently encountered by the practitioner in Discriminant Analysis is how to select the best variables. In mixed discriminant analysis (MDA), i.e., discriminant analysis with both continuous and discrete variables, the problem is more difficult because of the different nature of the variables. Various methods have been proposed in recent years for selecting variables in MDA. Here we use two versions of a generalized Mahalanobis distance between populations based on the Kullback-Leibler divergence for the first and on the Hellinger-Matusita distance for the second. Stopping rules are established from distributional results.

1.2 Generalisation of the Mahalanobis Distance

1.2.1 Introduction

Following Krzanowski (1983) the various distances proposed in the literature can be broadly classified in two categories:

1. Measures based on ideas from information theory (like Kullback-Leibler measures of information for example)
2. Measures related to Bhattacharya's measure of affinity (like Matusita's distance for example)

A review of these distance measures can be found, for example, in Adhikari and Joshi (1956).

Mixture of continuous and discrete variables is frequently encountered in discriminant analysis. The location model (Olkin and Tate, 1961; Krzanowski, 1990) is one possible way to deal with these data. Gower (1966) proposed a formula for converting similarity to distance. Since this transformation corresponds to the transformation of Bhattacharya's measure of affinity to Matusita's distance, Krzanowski (1983) studied the properties of Matusita's distance in the framework of the location model. Since no distributional properties were obtained, Krzanowski (1984), proposed to use Monte Carlo procedures to obtain percentage points. This distance was also proposed as a tool of selection of variables (Krzanowski, 1983). Distributional results for Matusita will be presented in Section 1.2.3. At first we present another generalization of the Mahalanobis distance, J , based on the Kullback-Leibler divergence.

One of the aims of discriminant analysis is the allocation of unknown entities to populations that are known *a priori*. A preliminary matter for consideration before an outright or probabilistic allocation is made for an unclassified entity X is to test the assumption that X belongs to one of the predefined groups π_i ($i = 1, 2, \dots, n$). One way of approaching this question is to test if the smallest distance between X and π_i is null or not. Most of the results were obtained in the case of linear discriminant analysis where the probability distribution function of the populations is assumed to be normal and with a common variance-covariance matrix Σ (McLachlan, 1992). Generally, the squared Mahalanobis distance is computed between X and each population π_i . X will be assessed as atypical if the smallest distance is bigger than a given threshold. Formally a preliminary test is of the form:

$$H_0 : \min_i d(X, \pi_i) = 0 \quad \text{versus} \quad H_1 : \min_i d(X, \pi_i) > 0$$

In practical case, the assumption of normality can be unrealistic. For example in taxonomy or in medicine, discrete and continuous measurements are taken. We propose a preliminary test to the general parametric case

1.2.2 Kullback–Leibler Divergence

The idea of using distance to discriminate between population using both continuous and categorical variables was studied by various authors, see Cuadras (1989), Morales, Pardo and Zografos (1998), Nakanishi (1996), Núñez, Villarroya and Oller (2003). We generalise the Mahalanobis distance using the divergence defined by Kullback–Leibler (Kullback, 1959) between two generalised probability densities $f_1(X)$ and $f_2(X)$:

$$\begin{aligned} J &= J\{f_1(X); f_2(X)\} \\ &= \int \{f_1(X) - f_2(X)\} \log \frac{f_1(X)}{f_2(X)} d\lambda \end{aligned}$$

where λ , μ_1 and μ_2 are three probability measures absolutely continuous with respect to each other and f_i is the Radon–Nikodym derivative of μ_i with respect to λ .

Except the triangular inequality, the Kullback–Leibler distance has the properties of a distance. Moreover, if f_1 and f_2 are multivariate normal distributions with common variance-covariance matrix then $J(f_1; f_2)$ is equal to the Mahalanobis distance.

Application to the Location Model

Suppose that g continuous variables $X = (X_1, \dots, X_g)^\top$ and d discrete variables $Y = (Y_1, \dots, Y_d)^\top$ are measured on each unit and that the units are drawn from the population π_1 or the population π_2 .

Moreover suppose that the condition of the location model (Krzanowski, 1990) holds. This means that:

- The d discrete variables define a multinomial vector Z containing c possible states. The probability of observing state m in the population π_i is:

$$p_{im} > 0 \quad (m = 1, \dots, c) \quad \text{and} \quad \sum_{m=1}^c p_{im} = 1, \quad (i = 1, 2)$$

- Conditionally on $Z = m$ and π_i , the q continuous variables X follow a multivariate normal distribution with mean $\mu_i^{(m)}$, variance–covariance matrix $\Sigma_i^{(m)}$ and density:

$$f_{i,m}(X) = f(X | Z = m, \pi_i)$$

- For the sake of simplicity, we assume $\Sigma_1^{(m)} = \Sigma_2^{(m)} = \Sigma$.

Since the aim is to compute the distance between π_1 and π_2 on the basis of the measurement made on X and Z , the joint density of X and Z given π_i is needed:

$$\begin{aligned} f_i(x, z) &= \sum_{m=1}^c f_{i,m}(x) p(Z = m | \pi_i) \mathbf{I}(z = m) \\ &= \sum_{m=1}^c f_{i,m}(x) p_{im} \mathbf{I}(z = m) \end{aligned}$$

This model was extended by some authors. Liu and Rubin (1998) relaxed the normality assumption. Bedrick, Lapidus and Powell (2000) considered the inverse conditioning and end up with a probit model and de Leon and Carrière (2004) generalize the Krzanowski and Bedrick approach.

PROPOSITION 1.1 By applying the Kullback–Leibler measure of distance to the location model, we obtain:

$$J = J_1 + J_2 \tag{1.1}$$

with

$$J_1 = \sum_m (p_{1m} - p_{2m}) \log \frac{p_{1m}}{p_{2m}}$$

and

$$J_2 = \frac{1}{2} \sum_m (p_{1m} + p_{2m}) (\mu_1^{(m)} - \mu_2^{(m)})^\top \Sigma^{-1} (\mu_1^{(m)} - \mu_2^{(m)})$$

The proof is straightforward.

Remark: This expression is meaningless if $p_{im} = 0$.

COROLLARY 1.1 *If the continuous variables are independent of the discrete variables then:*

$$\mu_1^{(m)} = \mu_1 \quad \text{and} \quad \mu_2^{(m)} = \mu_2 \quad \text{for all } m$$

and

$$J = \sum_m (p_{1m} - p_{2m}) \log \frac{p_{1m}}{p_{2m}} + (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)$$

which means that the Kullback-Leibler distance is equal to the sum of the contribution of the continuous and the discrete variables. This result is logical since J_1 represents the information based on Z , and J_2 the information based on X knowing Z .

Asymptotic Distribution of the Kullback-Leibler Distance in the Location Model

Generally the p_{im} , μ_{im} and Σ are unknown and have to be estimated from a sample using a model. Consider that we have two samples of size n_1 and n_2 respectively available from the population π_1 and π_2 and let n_{im} be the number of individuals, in the sample drawn from π_i , occupying the state m of the multinomial variable Z . In the model, there are two kinds of parameters: those which depend on the populations, and noisy parameters which are independent from the populations. They can be considered as noisy parameters since this category of parameters is not involved in the distance J . For example, if the mean is modelled with an analysis of variance model:

$$\mu_{im} = \mu + \alpha_i + \beta_m$$

where α is the population effect and β the discrete state effect. The expression of the distance is:

$$\mu_{1m} - \mu_{2m} = \alpha_1 - \alpha_2$$

So the β_m can be considered to be noisy parameters since they are not involved in the distance.

Let p be the vector of probability associated to the multinomial state of Z then

$$\hat{p} = p(\hat{\eta}) \tag{1.2}$$

where $\eta = (\eta_a, \eta_{ib})$; η_a is the set of noisy parameters and η_{ib} is the set of parameters used to discriminate between two populations.

Let r be the cardinal of η_{ib} . In the case of the location model, the p_{im} are generally estimated through a log-linear model.

Let μ be the vector of the mean of the continuous variables for the different states of Z then:

$$\hat{\mu} = \mu(\hat{\xi}) \quad (1.3)$$

where $\xi = (\xi_a, \xi_{ib})$; ξ_a is the set of noisy parameters and ξ_{ib} is the set of parameters used to discriminate between two populations.

Let s be the cardinal of ξ_{ib} . In the case of the location model, the μ_{im} are generally estimated through an analysis of variance model. Asparoukhov and Krzanowski (2000) also studied the smoothing of the location model parameters.

The aim of this section is to study the distributional property of both parts of the distance to obtain a test and a confidence interval for the classical hypothesis. Formally the following hypothesis are tested:

$$\begin{array}{lll} H_{01} : J_1 = 0 & \text{versus} & H_{11} : J_1 > 0 \\ H_{02} : J_2 = 0 & \text{versus} & H_{12} : J_2 > 0 \\ H_0 : J = 0 \quad (H_{01} \cap H_{02}) & \text{versus} & H_1 : J > 0 \quad (H_{11} \cup H_{12}) \end{array}$$

Asymptotic Results

Let $\theta_i = (\eta_a, \xi_a, \eta_{ib}, \xi_{ib}) = (\theta_a, \theta_{ib})$ for $i = 1, 2$ where $\eta_a, \xi_a, \eta_{ib}, \xi_{ib}$ are defined in (1.2) and (1.3). The following regularity conditions are assumed:

- θ_i is a point of the parameter space Θ , which is assumed to be an open convex set in a $(r + s)$ -dimensional Euclidean space.
- $f(x, \theta_i)$ has continuous second-order partial derivatives with respect to the θ_i 's in Θ ,
- $\hat{\theta}_i$ is the maximum likelihood estimator of θ_i
- For all $\theta_i \in \Theta$,

$$\int \frac{\partial f(x, \theta_i)}{\partial \theta_i} d\lambda(x) = \int \frac{\partial^2 f(x, \theta_i)}{\partial^2 \theta_i} d\lambda(x) = 0 \quad i = 1, 2$$

- The integrals

$$c(\theta_i) = \int \left\{ \frac{\partial \log f(x, \theta_i)}{\partial \theta_i} \right\}^2 f(x, \theta_i) d\lambda(x) \quad i = 1, 2$$

are positive and finite for all $\theta_i \in \Theta$.

It is obvious that the location model satisfies these conditions. Let $\hat{J} = J(\hat{\theta})$ be an estimator of J .

PROPOSITION 1.2 Under $H_0: \theta_1 = \theta_2 = \theta_0$, when $n_1 \rightarrow \infty$, $n_2 \rightarrow \infty$ and $\frac{n_1}{n_2} \rightarrow u$:

$$\frac{n_1 n_2}{n_1 + n_2} \hat{J} \sim \chi^2(r + s) \quad (1.4)$$

where r and s are the dimension of the space generated by η_{ib} and ξ_{ib}

Proof:

$$\hat{J} = \int \left\{ f(x, \hat{\theta}_1) - f(x, \hat{\theta}_2) \right\} \log \left\{ \frac{f(x, \hat{\theta}_1)}{f(x, \hat{\theta}_2)} \right\} d\lambda(x) \quad (1.5)$$

Since $p_{im} > 0$, the regularity conditions are satisfied. Therefore, Under $H_0: \theta_1 = \theta_2 = \theta_0$ a Taylor expansion of first order of $f(x, \hat{\theta}_1)$ and $f(x, \hat{\theta}_2)$ at the neighbourhood of θ_0 can be used:

$$\begin{aligned} \hat{J} &= J + (\hat{\theta}_1 - \theta_1)^\top \frac{\partial J}{\partial \theta_1} + (\hat{\theta}_2 - \theta_2)^\top \frac{\partial J}{\partial \theta_2} \\ &\quad + \frac{1}{2} (\hat{\theta}_1 - \theta_1)^\top \frac{\partial^2 J}{\partial \theta_1^2} (\hat{\theta}_1 - \theta_1) + \frac{1}{2} (\hat{\theta}_2 - \theta_2)^\top \frac{\partial^2 J}{\partial \theta_2^2} (\hat{\theta}_2 - \theta_2) \\ &\quad + (\hat{\theta}_2 - \theta_2)^\top \frac{\partial^2 J}{\partial \theta_1 \partial \theta_2} (\hat{\theta}_1 - \theta_1) + \sigma(\hat{\theta}_1 - \theta_1) + \sigma(\hat{\theta}_2 - \theta_2) \end{aligned}$$

Under H_0 :

$$\frac{\partial J}{\partial \theta_1} = \int \left[\frac{\partial f(x, \theta_1)}{\partial \theta_1} \log \left\{ \frac{f(x, \theta_1)}{f(x, \theta_2)} \right\} - \frac{\partial f(x, \theta_1)}{\partial \theta_1} \frac{f(x, \theta_2)}{f(x, \theta_1)} \right] d\lambda(x) = 0$$

since $\theta_1 = \theta_2 = \theta_0$ and $\int \frac{\partial f(x, \theta_1)}{\partial \theta_1} = 0$. For the same reason $\frac{\partial J}{\partial \theta_2} = 0$

For all $i, j = 1, 2$:

$$\begin{aligned} \frac{\partial^2 J}{\partial \theta_i \partial \theta_j} &= (\hat{\theta}_i - \theta_i)^\top \int \frac{f''^2(x, \theta_0)}{f(x, \theta_0)} d\lambda(x) (\hat{\theta}_j - \theta_j) \\ &= (\hat{\theta}_i - \theta_i)^\top I(\theta_0) (\hat{\theta}_j - \theta_j) \end{aligned}$$

where $I(\theta_0)$ represents the information matrix of Fisher.

Asymptotically, under H_0 , (1.5) becomes:

$$\begin{aligned} \hat{j} &= \frac{1}{2}(\hat{\theta}_1 - \theta_0)^\top I(\theta_0)(\hat{\theta}_1 - \theta_0) + \frac{1}{2}(\hat{\theta}_2 - \theta_0)^\top I(\theta_0)(\hat{\theta}_2 - \theta_0) \\ &\quad + (\hat{\theta}_1 - \theta_0)^\top I(\theta_0)(\hat{\theta}_2 - \theta_0) \\ &= (\hat{\theta}_1 - \hat{\theta}_2)^\top I(\theta_0)(\hat{\theta}_1 - \hat{\theta}_2) \end{aligned}$$

Since $\hat{\theta}_i$ is the maximum likelihood estimator of θ_0 (Rao, 1973):
 $\sqrt{n_i}(\hat{\theta}_i - \theta_0) \sim N_p\{0, I^{-1}(\theta_0)\}$ ($i = 1, 2$) Then:

$$\begin{aligned} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\hat{\theta}_1 - \theta_0) &\sim N_p\left\{0, \frac{1}{1+u} I^{-1}(\theta_0)\right\} \\ \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\hat{\theta}_2 - \theta_0) &\sim N_p\left\{0, \frac{u}{1+u} I^{-1}(\theta_0)\right\} \end{aligned}$$

Then

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} I(\theta_0)^{\frac{1}{2}} (\hat{\theta}_1 - \hat{\theta}_2) \sim N_p(0, 1)$$

Finally,

$$\frac{n_1 n_2}{n_1 + n_2} (\hat{\theta}_1 - \hat{\theta}_2)^\top I(\theta_0) (\hat{\theta}_1 - \hat{\theta}_2) \sim \chi^2(r + s)$$

COROLLARY 1.2 Under H_{01} :

$$\frac{n_1 n_2}{n_1 + n_2} \hat{J}_1 \sim \chi^2(r) \quad \text{when } n_1 \rightarrow \infty, n_2 \rightarrow \infty \text{ and } \frac{n_1}{n_2} \rightarrow u$$

Proof: It is enough to apply the proposition 1.2 with $q = 0$, which means the absence of continuous variables.

PROPOSITION 1.3 Under H_{02} :

$$\frac{n_1 n_2}{n_1 + n_2} \hat{J}_2 \sim \chi^2(s) \quad \text{when } n_1 \rightarrow \infty, n_2 \rightarrow \infty \text{ and } \frac{n_1}{n_2} \rightarrow u$$

Proof: The proof is very similar to the proof of the proposition 1.2.

1.2.3 Asymptotic Distribution of Matusita Distance

Krzanowski (1983) used Bhattacharya's affinity measure:

$$\rho = \int f^{\frac{1}{2}}(x, \theta_1) f^{\frac{1}{2}}(x, \theta_2) d\lambda(x)$$

to define the distance:

$$\begin{aligned}\Delta &= \int \left\{ f^{\frac{1}{2}}(x, \theta_1) - f^{\frac{1}{2}}(x, \theta_2) \right\}^2 d\lambda(x) \\ &= 2 - 2\rho\end{aligned}$$

This distance is also known as the Hellinger distance. In the location model context Krzanowski has obtained:

$$K = 2 - 2 \sum_m (p_{1m} p_{2m})^{\frac{1}{2}} \exp\left\{-\frac{1}{8}(\mu_{1,m} - \mu_{2,m})^\top \Sigma^{-1}(\mu_{1,m} - \mu_{2,m})\right\}$$

Let $\theta_i = (\eta_a, \xi_a, \eta_{bi}, \xi_{bi}) = (\theta_a, \theta_{bi})$ for $i = 1, 2$. Under $H_0 = (\theta_1 = \theta_2)$, we have $\xi_{bi} = 0$ and $\eta_{bi} = 0$ for $i = 1, 2$.

Under the usual regularity conditions, we prove the following result:

PROPOSITION 1.4 *Let $u \in]0, 1[$, $\hat{K} = K(\hat{\theta}_1, \hat{\theta}_2)$ with*

$$\hat{K} = 2 - 2 \sum_m (\hat{p}_{1m} \hat{p}_{2m})^{\frac{1}{2}} \exp\left\{-\frac{1}{8}(\hat{\mu}_{1,m} - \hat{\mu}_{2,m})^\top \hat{\Sigma}^{-1}(\hat{\mu}_{1,m} - \hat{\mu}_{2,m})\right\}$$

Assume that $H_0: \theta_1 = \theta_2 = \theta_0$ is true and that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent asymptotically efficient estimates of θ_0 . Then for $n_1 \rightarrow \infty$, $n_2 \rightarrow \infty$, $n_1/n_2 \rightarrow u$

$$\frac{4n_1 n_2}{(n_1 + n_2)} K(\hat{\theta}_1, \hat{\theta}_2) \sim \chi^2(r + s)$$

Proof

Under $H_0: \theta_1 = \theta_2 = \theta_0$, we obtain:

$$K(\theta_0) = 0$$

$$\frac{\partial K}{\partial \theta_1} = \frac{\partial K}{\partial \theta_2} = 0$$

and

$$\frac{\partial^2 K}{\partial \theta_1^2} = \frac{\partial^2 K}{\partial \theta_2^2} = -\frac{\partial^2 K}{\partial \theta_1 \partial \theta_2} = \frac{1}{2} \int \frac{f'^2(x, \theta_0)}{f(x, \theta_0)} d\lambda(x) = \frac{1}{2} I(\theta_0)$$

where $I(\theta_0)$ is the information matrix of Fisher. Under usual regularity conditions (Bar-Hen and Daudin, 1995), the Taylor expansion of the affinity

at the neighborhood of θ_0 can be derived and using the previous result we have, under H_0 :

$$K(\hat{\theta}_1, \hat{\theta}_2) \approx \frac{1}{4}(\hat{\theta}_1 - \hat{\theta}_2)^\top I(\theta_0)(\hat{\theta}_1 - \hat{\theta}_2)$$

Since $\hat{\theta}_i$ are independent asymptotically efficient estimator of θ_0 , $n_i^{\frac{1}{2}}(\hat{\theta}_i - \theta_0) \sim N_p(0, I^{-1}(\theta_0))$ ($i = 1, 2$). Then:

$$\begin{aligned} \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{\frac{1}{2}}(\hat{\theta}_1 - \theta_0) &\sim N_p\left\{0, \frac{1}{1+u}I^{-1}(\theta_0)\right\} \\ \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{\frac{1}{2}}(\hat{\theta}_2 - \theta_0) &\sim N_p\left\{0, \frac{u}{1+u}I^{-1}(\theta_0)\right\} \end{aligned}$$

Then

$$\left(\frac{n_1 n_2}{n_1 + n_2}\right)^{\frac{1}{2}} I(\theta_0)^{\frac{1}{2}} (\hat{\theta}_1 - \hat{\theta}_2) \sim N_p(0, 1)$$

Additional results can be found in Bar-Hen and Daudin (1998).

1.2.4 Simulations

The level and the power of the test described in the previous section were evaluated through simulations. One continuous variable and two binary variables are considered. Hence the multinomial vector Z has 4 levels. The estimates of the means, the proportions and the variance are the maximum likelihood estimates. These estimates corresponds to saturated model and therefore the test of the distance has 7 degrees of freedom. It has to be noted that no correction factor for the case $p_{im} = 0$ and therefore empty cells are taken into account for the computation of the distance.

Four cases were studied:

1. no population effect for the discrete variables and no population effect for the continuous variables ($K = 0$);
2. no population effect for the discrete variables but a population effect for the continuous variables;
3. a population effect for the discrete variables but no population effect for the continuous variables;

4. a population effect for the discrete and the continuous variables.

For the continuous variables, the population effect is equal to the standard error:

$$\frac{\mu_{1,m} - \mu_{2,m}}{\sigma} = \begin{cases} 0 & \text{if population effect is present} \\ 1 & \text{if population effect is not present} \end{cases}$$

For the discrete variables:

$$\log \left(\frac{p_{1m}}{p_{2m}} \right) = \begin{cases} 0 & \text{if population effect is present} \\ 1 & \text{if population effect is not present} \end{cases}$$

Since the aim of these simulations is to estimate the rate of convergence of the asymptotic distributions, populations of size 20 and 100 were considered. This gives three new cases:

1. population π_1 of size 10 and population π_2 of size 10
2. population π_1 of size 30 and population π_2 of size 30
3. population π_1 of size 100 and population π_2 of size 100

There are 12 combinations of hypotheses and populations sizes. 1000 simulations were done for each combination. The table below presents the number of non-significant tests at the 5% level.

By using the property of the binomial distribution, one may expect to obtain $50 \pm 1.96 \times (1000 \times 0.5 \times 0.95)^{\frac{1}{2}} = 50 \pm 14$ tests to be non-significant if the null hypothesis is true.

From Table 1.1, we deduce that the level of the test is respected as soon as $n \geq 30$. This means 30/4 observations per cell. The power of the test tends to 1 but the convergence is slower for the discrete variables. This result is not surprising.

It has to be noted that these simulations are limited. The use of non-saturated model for the estimation of the parameters and the use of a correction factor for empty cell can probably alter the results.

1.3 Methods and Stopping Rules for Selecting Variables

As in the usual discriminant analysis with continuous variables, selection of variables is a problem of practical importance. In fact, in the location model

Table 1.1: Number of significant test at the 5% level for the various hypotheses

population effect for		size of population		Hypothesis tested
discrete var.	continuous var.	π_1	π_2	$K = 0$
no	no	10	10	68
no	no	30	30	60
no	no	100	100	60
no	yes	10	10	251
no	yes	30	30	798
no	yes	100	100	1000
yes	no	10	10	144
yes	no	30	30	255
yes	no	100	100	711
yes	yes	10	10	344
yes	yes	30	30	872
yes	yes	100	100	1000

context, the question is more precisely "which terms and which continuous variables must be included in the model?" where the models concerned are log-linear and MANOVA. Interest in this topic has been shown regularly since the paper published by Vlachnonikolis and Marriot (1982). Krzanowski (1983) used a Matusita-Hellinger distance between the populations, Daudin (1986) used a modified AIC method and Krusinska (1989), Krusinska (1990) used several methods based on the percentage of misclassification, Hotelling's T^2 and graphical models.

Based on Hellinger distance, Krzanowski (1983) proposed the use of a distance K to determine the most discriminative variables.

Our asymptotic results allow us to propose stopping rules based on the P -value of the test of $J = 0$ or $K = 0$. These two methods were then compared with a third, based on the Akaike Information Criterion (AIC) described by Daudin (1986): classically, AIC penalize the likelihood by the number of parameters. A direct use of AIC on MANOVA models (described in Section 1.2.2) will lead to noncomparable log-likelihood. Daudin (1986) proposed to eliminate the noisy parameters (noted β_m) and to penalize the log-likelihood by the number of parameters related to the population effect. It permits to judge whether the log-likelihood and the increase of AIC is only due to population factor terms in the ANOVA model and is not coming from noisy parameters.

Krzanowski (1983) used the distance K to select variables. It should be noted that \hat{K} increases when the location model contains more variables without guaranteeing that this increase is effective: it is therefore necessary to discount any slight increase that may be caused by chance. We propose to include a new discriminant variable or a new term in the location model if it increases the evidence that H_0 ($K = 0$) is false as measured by the P -value of the test of the null hypothesis, using the asymptotic distribution of \hat{K} .

It would be interesting to test whether the increase of K due to a new term in the model is positive. Unfortunately when K is positive (H_0 false) the asymptotic distribution of the increase in \hat{K} due to a new term is not easily tractable under the hypothesis that the new parameter is null.

An alternative criterion is an Akaike-like one: $K - AIC = 4 \frac{n_1 n_2}{n_1 + n_2} \hat{K} - 2(r + s)$. According to this method, the best model is that which maximizes $K - AIC$.

It is also possible to use \hat{J} with the same methods: we can use the P -value of the chi-square test of $J = 0$ or alternatively $J - AIC = \frac{n_1 n_2}{n_1 + n_2} \hat{J} - 2(r + s)$

Based on simulations, Daudin and Bar-Hen (1999) showed that all three competing methods (two distances and Daudin-AIC) gave good overall performances (nearly 85% correct selection). The K -method has weak power with discrete variables when sample sizes are small but is a good choice when a simple model is requested. The J -method possesses an interesting decomposition property of $J = J_1 + J_2$ between the discrete and continuous variables. The K -AIC and J -AIC methods select models that have more parameters than the P -value methods. For distance, the K -AIC method may be used with small samples, but the J -AIC method is not interesting for it increases the overparametrization of the $J - P$ method. The Daudin-AIC method gives good overall performance with a known tendency toward overparametrization.

1.4 Reject Option

1.4.1 Distributional Result

Since the aim is to test the atypicality of X , we have to derive the distribution of the estimate of the divergence J between X and π_i under the hypothesis $J(X, \pi_i) > 0$. We don't make assumptions about the distribution of the populations but the same regularity conditions as before are assumed. Bar-Hen and Daudin (1997) and Bar-Hen (2001) considered the reject option for the case of normal populations.