STATISTICAL MODELING BY STOCHASTIC COMPLEXITY

By

Guoqi Qian

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
AUGUST 12, 1994

ISBN 0-315-98927-0

Name GuoQi QIAN

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Statistics
SUBJECT TERM

0463
SUBJECT CODE

UMI

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS
- Architecture 0729
- Art History 0377
- Cinema 0900
- Dance 0378
- Fine Arts 0357
- Information Science 0723
- Journalism 0391
- Library Science 0399
- Mass Communications 0708
- Music 0413
- Speech Communication 0459
- Theater 0465

EDUCATION
- General 0515
- Administration 0514
- Adult and Continuing 0516
- Agricultural 0517
- Art 0273
- Bilingual and Multicultural 0282
- Business 0688
- Community College 0275
- Curriculum and Instruction 0727
- Early Childhood 0518
- Elementary 0524
- Finance 0277
- Guidance and Counseling 0519
- Health 0680
- Higher 0745
- History of 0520
- Home Economics 0278
- Industrial 0521
- Language and Literature 0279
- Mathematics 0280
- Music 0522
- Philosophy of 0998
- Physical 0523
- Psychology 0525
- Reading 0535
- Religious 0527
- Sciences 0714
- Secondary 0533
- Social Sciences 0534
- Sociology of 0340
- Special 0529
- Teacher Training 0530
- Technology 0710
- Tests and Measurements 0288
- Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS
- Language
  - General 0679
  - Ancient 0289
  - Linguistics 0290
  - Modern 0291
- Literature
  - General 0401
  - Classical 0294
  - Comparative 0295
  - Medieval 0297
  - Modern 0298
  - African 0316
  - American 0591
  - Asian 0305
  - Canadian (English) 0352
  - Canadian (French) 0355
  - English 0593
  - Germanic 0311
  - Latin American 0312
  - Middle Eastern 0315
  - Romance 0313
  - Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY
- Philosophy 0422
- Religion
  - General 0318
  - Biblical Studies 0321
  - Clergy 0319
  - History of 0320
  - Philosophy of 0322
- Theology 0469

SOCIAL SCIENCES
- American Studies 0323
- Anthropology
  - Archaeology 0324
  - Cultural 0326
  - Physical 0327
- Business Administration
  - General 0310
  - Accounting 0272
  - Banking 0770
  - Management 0454
  - Marketing 0338
- Canadian Studies 0385
- Economics
  - General 0501
  - Agricultural 0503
  - Commerce Business 0505
  - Finance 0508
  - History 0509
  - Labor 0510
  - Theory 0511
- Folklore 0358
- Geography 0366
- Gerontology 0351
- History
  - General 0578
  - Ancient 0579
  - Medieval 0581
  - Modern 0582
  - Black 0328
  - African 0331
  - Asia, Australia and Oceania 0332
  - Canadian 0334
  - European 0335
  - Latin American 0336
  - Middle Eastern 0333
  - United States 0337
- History of Science 0585
- Law 0398
- Political Science
  - General 0615
  - International Law and Relations 0616
  - Public Administration 0617
- Recreation 0814
- Social Work 0452
- Sociology
  - General 0626
  - Criminology and Penology 0627
  - Demography 0938
  - Ethnic and Racial Studies 0631
  - Individual and Family Studies 0628
  - Industrial and Labor Relations 0629
  - Public and Social Welfare 0630
  - Social Structure and Development 0700
  - Theory and Methods 0344
- Transportation 0709
- Urban and Regional Planning 0999
- Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES
- Agriculture
  - General 0473
  - Agronomy 0285
  - Animal Culture and Nutrition 0475
  - Animal Pathology 0476
  - Food Science and Technology 0359
  - Forestry and Wildlife 0478
  - Plant Culture 0479
  - Plant Pathology 0480
  - Plant Physiology 0817
  - Range Management 0777
  - Wood Technology 0746
- Biology
  - General 0306
  - Anatomy 0287
  - Biostatistics 0308
  - Botany 0309
  - Cell 0379
  - Ecology 0329
  - Entomology 0353
  - Genetics 0369
  - Limnology 0793
  - Microbiology 0410
  - Molecular 0307
  - Neuroscience 0317
  - Oceanography 0416
  - Physiology 0433
  - Radiation 0821
  - Veterinary Science 0778
  - Zoology 0472
- Biophysics
  - General 0786
  - Medical 0760

EARTH SCIENCES
- Biogeochemistry 0425
- Geochemistry 0996
- Geodesy 0370
- Geology 0372
- Geophysics 0373
- Hydrology 0388
- Mineralogy 0411
- Paleobotany 0345
- Paleoecology 0426
- Paleontology 0418
- Paleozoology 0985
- Palynology 0427
- Physical Geography 0368
- Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES
- Environmental Sciences 0768
- Health Sciences
  - General 0566
  - Audiology 0300
  - Chemotherapy 0992
  - Dentistry 0567
  - Education 0350
  - Hospital Management 0769
  - Human Development 0758
  - Immunology 0982
  - Medicine and Surgery 0564
  - Mental Health 0347
  - Nursing 0569
  - Nutrition 0570
  - Obstetrics and Gynecology 0380
  - Occupational Health and Therapy 0354
  - Ophthalmology 0381
  - Pathology 0571
  - Pharmacology 0419
  - Pharmacy 0572
  - Physical Therapy 0382
  - Public Health 0573
  - Radiology 0574
  - Recreation 0575
  - Speech Pathology 0460
  - Toxicology 0383
- Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences
- Chemistry
  - General 0485
  - Agricultural 0749
  - Analytical 0486
  - Biochemistry 0487
  - Inorganic 0488
  - Nuclear 0738
  - Organic 0490
  - Pharmaceutical 0491
  - Physical 0494
  - Polymer 0495
  - Radiation 0754
- Mathematics 0405
- Physics
  - General 0605
  - Acoustics 0986
  - Astronomy and Astrophysics 0606
  - Atmospheric Science 0608
  - Atomic 0748
  - Electronics and Electricity 0607
  - Elementary Particles and High Energy 0798
  - Fluid and Plasma 0759
  - Molecular 0609
  - Nuclear 0610
  - Optics 0752
  - Radiation 0756
  - Solid State 0611
- Statistics 0463

Applied Sciences
- Applied Mechanics 0346
- Computer Science 0984
- Engineering
  - General 0537
  - Aerospace 0538
  - Agricultural 0539
  - Automotive 0540
  - Biomedical 0541
  - Chemical 0542
  - Civil 0543
  - Electronics and Electrical 0544
  - Heat and Thermodynamics 0348
  - Hydraulic 0545
  - Industrial 0546
  - Marine 0547
  - Materials Science 0794
  - Mechanical 0548
  - Metallurgy 0743
  - Mining 0551
  - Nuclear 0552
  - Packaging 0549
  - Petroleum 0765
  - Sanitary and Municipal 0554
  - System Science 0790
- Geotechnology 0428
- Operations Research 0796
- Plastics Technology 0795
- Textile Technology 0994

PSYCHOLOGY
- General 0621
- Behavioral 0384
- Clinical 0622
- Developmental 0620
- Experimental 0623
- Industrial 0624
- Personality 0625
- Physiological 0989
- Psychobiology 0349
- Psychometrics 0632
- Social 0451

To My Wife Hong Yao and Our Family

Contents

List of Tables viii

Abstract ix

Acknowledgements x

1 General Introduction 1

1.1 Introduction 1

1.2 Modeling and Coding of Data 2

1.3 Coding of Integers 6

1.4 Complexity in a Coding System 10

1.5 Two-part Codes 12

1.6 Stochastic Complexity 19

1.7 Predictive Coding 22

1.8 Principle of Minimum Description Length 26

2 Principal Components Selection by the Criterion of the Minimum Mean Difference of Complexity 31

2.1 Introduction 31

2.2 Selection Principle for model $\{P(\underline{X}|\,\theta)\}$ 34

2.3 Principal Components in Normality Case 37

2.4 Validation of the Principal Components 39

2.5 "ε-contaminated" Normal Random Vectors 44

3 Generalized Linear Model Selection by Predictive Least Quasi-deviance Criterion **53**

3.1 Introduction 53

3.2 Model Selection and the "Honest" Predictive Error 56

3.3 The Predictive Least Quasi-deviance Criterion 60

3.3.1 Main Results 60

3.3.2 Remarks on Some Conditions of Theorem 3.3.1 and 3.3.2 . . . 63

3.4 An Approximate PLQD and A Monte Carlo PLQD 64

3.5 A Simulation Study 66

3.6 Proofs of Theorems 3.3.1 and 3.3.2 69

4 On Stochastic Complexity Estimation — A Decision Theoretic Approach **78**

4.1 Introduction 78

4.2 Complexity Decision Rule 83

4.3 Application to Parametric Families 90

4.4 Minimum Expected Risk and Admissibility 93

4.5 Completeness 97

5 Stochastic Complexity in Histograms and Testing Homogeneity **103**

5.1 Introduction 103

5.2 Data Compression for Optimal Information Description 105

5.3 Hypothesis Testing for Homogeneity 115

5.3.1 Background 115

5.3.2 The Test Procedure 117

5.3.3 Two Examples 126

5.3.4 Simulation Studies 128

5.4 Proofs of the Theorems 140

6 Concluding Remarks **157**

6.1 Summary 157

6.2 Future Research 158

A Programs for Chapter 3 **160**

B Programs for Chapter 5 **182**

Bibliography **199**

List of Tables

2.1 47
2.2 Principal Components based on model (2.5.1) 51
2.3 Principal Components based on the Sample Covariance Matrix 51
2.4 Principal Components based on the MVE Covariance Matrix Estimate 52
2.5 Principal Components based on the Weighted Covariance Matrix Estimate 52

3.1 *The Values of x_{ki} in Example 3.5.1* 74
3.2 Probabilities (Based on 1000 Repetitions) of Selecting Each Model . . 75
3 3 Continued to Table 3.2 76
3.4 *The Values of x_{ki} in Example 3.5.2* 76
3.5 Probabilities (Based on 1000 Repetitions) of Selecting Each Model . . 77
3.6 A comparison of site preferences of two species of lizards, *grahami* and *opalinus* 77

Abstract

This thesis is a study of several statistical modeling problems by stochastic complexity.

At first, an index of predictive power, using the concept of complexity or minimum description length, is proposed as a criterion to select the principal components of a random vector distributed in a parametric family.

Then, we consider the problem of selecting a model with the best predictive ability in a class of generalized linear models. A predictive least quasi-deviance criterion is proposed to measure the predictive ability of a model. Some results concerning the consistency of this criterion are given. The method is also modified for finite sample applications.

Thirdly a density estimation based complexity decision rule is proposed, which uses the quality of these estimators to estimate the corresponding unknown element of the true probability density. The resulting complexity density decision procedure is shown to be admissible, to achieve the minimum expected risk, and to form a minimal complete class.

Fourthly a generalized histogram density estimator with unequal-width subintervals is used to find both optimal and predictive optimal description of a sample. Both optimal descriptions are expressed in terms of the stochastic complexity. Uniform, almost sure asymptotic expressions for both descriptions are given.

Finally, as an application of the stochastic complexity for optimal data description, a new test procedure for hypotheses of homogeneity is proposed. Some examples and simulation studies are further given to illustrate this test procedure.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. R. P. Gupta and Dr. George Gabor, for their invaluable advice and help during the preparation of this thesis. Dr. Gabor introduced me to the wonderland of stochastic complexity in which I found my major research interests. Fruitful discussions with him induced me to writing Chapter 5 and studying the data compression and the quantization of information sources. Dr. Gupta was a constant source of encouragement and support. His excellent guidance in multivariate statistics made it possible for me to complete the work presented in Chapter 2. His generous help was also far beyond the preparation of this thesis. Further appreciation is extended to my supervisors for their careful reading and helpful suggestions, which removed many obscurities and led to the considerable improvement of the whole thesis.

I would also like to thank Dr. Chris Field and Dr. Keith Thompson for their service on my final examination committee. Special thanks go to Dr. Field for lots of his help, both academic and non-academic, and his continuing encouragement during my years as a graduate student.

I am particularly grateful to the external examiner, Dr. Jorma Rissanen. His inspiration and comments and helpful discussions with him significantly benefited the writing of this thesis. He also sent his papers to me prior to their publication.

This research work received the financial support from Dalhousie University and the Killam Trustees through an Izaak Walton Killam Memorial Scholarship, which are gratefully acknowledged.

Finally I would like to express my deep appreciation to my parents and my wife

Hong Yao for their love and support. I am especially indebted to my wife Hong for her love, understanding and great encouragement. Many hours that I have spent on the preparation of the thesis could otherwise have been devoted to her.

Chapter 1

General Introduction

1.1 Introduction

In recent years a new general approach to problems of statistical inference, stochastic complexity, has been developed by Jorma Rissanen. This approach takes the point of view that any statistical model is merely a human attempt to describe or explain the truth in the system generating data; and that such models are to be assessed in terms of their success at this task. In the theory of stochastic complexity, the model assessment is conducted under the principle of minimum description length (MDL). To find the description length (or predictive description length, or stochastic complexity) of an employed model with a sequence of an observed data string, a prefix coding procedure is supplied to encode the data string into a sequence of binary digits in two steps. The first step is to encode the data string under the employed model while the second step is to give a codeword indicating how complex it needs to be to specify the employed model in the assumed model class. The resulting two-part code length gives a measurement indicating the success of describing or explaining the random structure in the observed data string. The stochastic complexity is an abstract notion giving the shortest required length for describing the data by using the models in the assumed model class (or model classes). It provides the rationale for the minimum description length principle, which was developed under the inspiration

of the algorithmic notion of information by Solomonoff (1964), Kolmogorov (1965), Chaitin (1975) and others.

The theory of stochastic complexity has a great potential in statistical analysis. It is well suited to statistical model selection, where it generalizes the maximum likelihood principle, the maximum entropy principle, Akaike's AIC and Jeffreys-Schwarz-BIC penalized log-likelihood criterion. This thesis studies several statistical modeling problems by applying the idea of stochastic complexity. It includes the principal components selection in multivariate analysis, generalized linear model selection, decision settlement of stochastic complexity estimation, nonparametric testing hypothesis of homogeneity and general nonparametric histogram density estimation. However, before the full display of the study, we will briefly introduce the theory of stochastic complexity in this chapter.

First we describe the connection between modeling and coding of data. Then we describe the coding of integers. As an important element in the development of stochastic complexity, the complexity in a coding system is demonstrated. Section 1.5 to Section 1.8 gives the main part of the theory of stochastic complexity, including the two-part codes, the stochastic complexity, the predictive coding and the minimum description length principle.

The materials in this chapter mainly come from the first three chapters of Rissanen's "Stochastic Complexity in Statistical Inquiry (1989)", and from his papers (1983,1986a,1987). In addition, we describe some of the latest development in this area.

1.2 Modeling and Coding of Data

When encountering a real world phenomenon it is often necessary and useful to understand it, to find out the pattern it follows, and then in turn to improve our understanding of it. This way of knowing the world might be accomplished by regarding the phenomenon studied as being generated from an unknown system, and

by describing the structure and behavior of the system.

To describe and analyze the system is not easy because what are generally available are only the observations about the system, not the whole system itself. We must collect measurements of various kinds, which we think give us information about the unknown system, and then try to piece them together to give us an understanding of its secrets. Based on this understanding we explain the observed phenomenon and possibly further give a prediction for the future. This procedure of finding a pattern in the observed data is called model building or modeling.

In information theory, description of a system car be made by the way of coding. Let A denote a finite or countable set called an alphabet. Write A^n for the set of all strings of length n — each string consists of elements of A — and $A^* = \bigcup_{n=0}^{\infty} A^n$ for their union. For convenience, A^0 consists of the empty string, written as λ. The system of study is usually referred to an information source $\{A, P\}$, which is defined by the alphabet A and a probability function P with domain A^* and range $[0, 1]$ such that $P(\lambda) = 1$. The definition of information source is so general that a great deal of flexibility of study is allowed. The observed measurements of the system are expressed in terms of a finite string $x = x_1, \cdots, x_n \in A^*$, called a message. The coding of the observed message is important for studying the complexity and properties of the information source.

A code C is a single valued mapping from A^* into B^*, the set of all finite binary strings. For a message x in A^*, $C(x)$ is also called a code without any confusion. Nothing essential is lost by restricting the code alphabet to be binary.

To write a code for any message $x = x_1, \cdots, x_n$ in A^*, it is enough to define a codeword for each element in A if we assume all the x_i's are generated independently. In this way a message x in A^* can be encoded into a binary string by replacing x_i's by their corresponding codewords and concatenating them together without any commas. It is desirable that any encoded message of A^* can also be decoded back instantaneously. The word "instantaneously" means that for any code of a message of A^*, we can decode to the point we have reached with no necessity of reading the

whole code first.

If the code C is instantaneously decodable, it is called a prefix code. To ensure C is a prefix code, the Kraft inequality must be hold, i.e.

$$\sum_{a \in A} 2^{-L(a)} \leq 1 \tag{1.2.1}$$

where $L(a) = |C(a)|$ denotes the length of the codeword for a. Conversely, if we are given a sequence of positive integers $n_0, n_1, \cdots, n_k$ satisfying the Kraft inequality $\sum_{i=0}^{k} 2^{-n_i} \leq 1$ (here k could tend to infinity), we are also able to construct a prefix code for each element of alphabet $\{0, \cdots, k\}$ with length defined by these integers. Therefore, the Kraft inequality is equivalent to the prefix property, Rissanen (1989, p. 23).

One of the main objectives with coding in information theory is to shorten the description of a long data string (message). The question arises of how to construct an optimal prefix code for an information source $\{A, P\}$. The formulation of the optimization problem can vary. However, it is related to the distribution P and the following inequality plays a fundamental role in answering this question.

Theorem 1.2.1 *Let A be a finite or countable set, and let P and Q be two distributions on A. Then*

$$-\sum_{a \in A} P(a) \log Q(a) \geq -\sum_{a \in A} P(a) \log P(a). \tag{1.2.2}$$

Moreover, the equality holds if and only if $Q(a) = P(a)$.

Here and thereafter in this thesis, the logarithm is base 2 unless otherwise indicated. The proof of this theorem can easily be completed by using Jensen's inequality and therefore is omitted.

Suppose C is a prefix code for A^*, i.e. (1.2.1) is true. Then we can define a distribution on A as follows,

$$Q(a) = \frac{2^{-L(a)}}{\sum_{x \in A} 2^{-L(x)}} \quad \text{for any } a \in A. \tag{1.2.3}$$

From Theorem 1.2.1 we have

$$\sum_{a\in A} P(a)L(a) + \log\left(\sum_{a\in A} 2^{-L(a)}\right) \geq -\sum_{a\in A} P(a)\log P(a) \tag{1.2.4}$$

The inequality (1.2.4) can be interpreted to mean that for any prefix code C, mean code length of C is bounded from below by the entropy

$$H(P) = -\sum_{a\in A} P(a)\log P(a). \tag{1.2.5}$$

This is the famous noiseless coding theorem due to Shannon (1948).

On the other hand, with the distribution P defined by the information source $\{A, P\}$, a prefix code for A can be constructed whose mean code length does not differ from the entropy by more than one bit. In Section 2.2.2 of Rissanen (1989) an elegant algorithm — due to Huffman — for constructing an optimal prefix code for A was given with code length as close as possible to $-\log P(a)$ for each $a \in A$. The perfect match for $-\log P(a)$ is not possible unless the probabilities of a_i's are integer powers of $1/2$.

Ignoring the difference of at most one bit, define $L_P(a) = -\log P(a)$ for $a \in A$, as a length function generated by P. Then we have the following results due to Dawid (1992).

Theorem 1.2.2 *For any information source $\{A, P\}$ with finite or countable alphabet and a prefix code C with length function L_C, we have for all $\varepsilon > 0$,*

$$P(L_C(a) \leq L_P(a) - \varepsilon) \leq 2^{-\varepsilon}. \tag{1.2.6}$$

If we use $x^n = x_1, x_2, \cdots, x_n$ to denote any message in $\{A, P\}$, then further $L_P(x^n) - L_C(x^n)$ is bounded above with P-probability 1 as $n \to \infty$.

Proof: Denote $E_\varepsilon = \{a \in A | L_C(a) \leq L_P(a) - \varepsilon\}$. Then for any $a \in E_\varepsilon$, it is easily shown that $P(a) \leq 2^{-\varepsilon} 2^{-L_C(a)}$. Summing over all elements of E_ε and applying the Kraft inequality gives $P(E_\varepsilon) \leq 2^{-\varepsilon} \sum_{a\in E_\varepsilon} 2^{-L_C(a)} \leq 2^{-\varepsilon}$ which is (1.2.6).

For the proof of the second part, we assume without loss of generality that C satisfies Kraft inequality with equality. Otherwise we could shorten the code length

by $-\log\left(\sum_{a\in A} 2^{-L_C(a)}\right)$ for each element $a \in A$. Define $U_n = 2^{L_P(x^n)-L_C(x^n)} = P_C(x^n)/P(x^n)$ where $P_C(x^n) = 2^{-L_C(x^n)}$ can be shown to be a distribution. It is readily shown that U_n is a non-negative martingale under P and hence is bounded above with P-probability 1 as $n \to \infty$ and so is $L_P(x^n) - L_C(x^n)$. □

Theorem 1.2.1 and Theorem 1.2.2 may be taken as establishing L_P as a length function for the optimal prefix code of the information source $\{A, P\}$. In particular, if we apply Theorem 1.2.2 to the encoding of long sequences of symbols, the per-symbol message length achieved by any prefix code can not improve on that given by L_P by more than a negligible amount, with arbitrarily high probability under P.

If we treat A^n, the set of all strings of length n in $\{A, P\}$, as a new alphabet, an extended information source $\{A^n, P^n\}$ is obtained, where P^n is defined by independence as assumed above. With these arguments we can construct an optimal prefix code for A^n with mean code length not differing from the entropy $H(P^n) = nH(P)$ by one bit.

So far we have discussed the coding which treats the symbol occurrences as independent only. In practical situations, the independence condition cannot always be guaranteed and for this reason, a powerful coding technique, the arithmetic coding, which is designed to do the coding for general discrete random process, stationary or not, was introduced. For detail, see Rissanen (1976, 1989), Rissanen and Mohiuddin (1989), and Rissanen and Langdon (1981).

1.3 Coding of Integers

In addition to encoding a message coming from an information source $\{A, P\}$, we also need to encode in a prefix manner, positive integers for which no distribution is given. There is an efficient prefix code due to Elias (1975) for the set of positive integers, which we describe below.

To understand the code construction, we start by encoding the integer n as its binary representation. Such a code cannot be a prefix code, because its length function

or its upper bound $\log(2n)$, does not satisfy the Kraft inequality. On the other hand, if the binary representation were followed by other binary symbols, as is usual in the case when we encode a set of integers, we would not be able to recognize where the representation ends. To overcome this difficulty, we supply the length l_1 of the binary representation of n as a preamble, the length l_2 of the binary representation of $l_1 - 1$ as another preamble, the length l_3 of the binary representation of $l_2 - 1$ as a third preamble, and so on, until the k-th step where $l_k = 2$. By this iteration, we obtain a monotone decreasing sequence of integers $n, l_1, l_2, l_3, \cdots, l_k$. Now we find the binary representations of $l_{k-1} - 1$, $l_{k-2} - 1$, $\cdots, l_1 - 1$, n and paste them together, and add a symbol 0 to the end to indicate that the preceding binary representation is for the integer n. By doing so, we construct a prefix code $w(n)$ for n. Some examples are:

$$w(1) = 0,\ w(2) = 10\ 0,\ w(3) = 11\ 0,\ w(4) = 10\ 100\ 0,\ w(7) = 10\ 111\ 0,$$
$$w(14) = 11\ 1110\ 0, w(15) = 11\ 1111\ 0,\ w(16) = 10\ 100\ 10000\ 0,$$
$$w(65651) = 10\ 100\ 10000\ 10011011100010011\ 0.$$

Here we insert some blanks in the codes for easier reading which, of course, are not needed to decode the number n. Note that l_k, the final length, is 2, so $l_{k-1} = 3$ or 4, and the binary representation of $l_{k-1} - 1$ is either 10 or 11. If $k = 1$ which implies $l_1 = 2$ or 1, then $n = 1$ or 2 or 3; the code of which is 0 or 100 or 110. For codes other than these three, we can decode it as follows. First decode the first two symbols in the code to the length l_{k-1}. Then using this information decode the next l_{k-1} symbols to get l_{k-2}, and so on, until decode the binary representation of n. For example, we decode 16 out of $w(16)$. We get 2 by decoding the first two symbols 10, this tells us to decode the next 3 symbols 100 which returns 4, so we need to decode the next 5 symbols 10000, which is 16, then we run against 0 which means 16 is n but not the length information.

It is apparent that the length function of this code is approximately

$$L(n) = \log^* n + \log c \tag{1.3.1}$$

where $\log^* = \log n + \log\log n + \cdots$ including only the non-negative terms, and c is

a constant such that the Kraft inequality holds. Among these values of the constant c, we can select one which satisfies the Kraft inequality with equality. Such a value of c is $c^* \approx 2.865064$, Leung-Yan-Cheong and Cover (1978) and Rissanen (1983). From Section 1.2 we know there exists a prefix code with length function $L^*(n) = \log^* n + \log c^*$ and further from Bentley and Yao (1976) we know that any monotone non-decreasing length function $L(n)$ of positive integers, which satisfies the Kraft inequality, must equal or exceed $L^*(n) - 2k^*(n)$ infinitely often, where $k^*(n)$ denotes the number of terms in $\log^*(n)$.

Define $Q^*(n) = 2^{-L^*(n)}$, Q^* is a distribution on the set of positive integers and by (1.3.1)

$$Q^*(n) = (c^* n \log n \log\log n \cdots)^{-1}. \tag{1.3.2}$$

Rissanen calls $Q^*(n)$ the universal prior for the positive integers. This prior can be extended to all non-negative integers by defining $Q^*(0) = 1/2$ and replacing c^* by $2c^*$. To extend this distribution to the set of all integers, add one to $L^*(n)$ and define $Q^*(-n) = Q^*(n)$.

Q^* has the following optimum property, Rissanen (1983).

Theorem 1.3.1 *For any distribution $P(n)$ for the positive integers such that*

$$(i) \quad P(n) \geq P(n+1), \qquad n > M, \text{ for some } M \tag{1.3.3}$$

$$(ii) \qquad -\textstyle\sum_{n\geq 1} P(n)\log P(n) = \infty, \tag{1.3.4}$$

the following holds

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} P(n)L^*(n)}{-\sum_{n=1}^{N} P(n)\log P(n)} = 1. \tag{1.3.5}$$

Since it follows from Theorem 1.2.1 that the limit can not be smaller than unity, we conclude that, if we encode large integers with the code length $L^*(n)$, we can do no better even if with a distribution $P(n)$ with which to design the code. Hence, $L^*(n)$ can be taken to be just about the ideal code length for large positive integers.

Besides encoding one integer in a prefix manner, frequently there is a need to encode a set of them. If the integers in this set are completely independent in the sense that no one is affected by the others, then we only need paste their individual prefix codes together to encode them. However, it is often the case that the integers are of the same order of magnitude, a fact which we can take advantage of. Consider, then, a set of integers $n_1, n_2, \cdots, n_m$, of which, say, m_+ are non-negative. A prefix code of them can be constructed with about

$$L(n_1, \cdots, n_m) = L^*(n) + \log \frac{(n+m)!}{n!(m-1)!} + \log \frac{(m+1)!}{m_+!(m-m_+)!} \tag{1.3.6}$$

bits, where $n = \sum_{i=1}^{m} |n_i|$, Rissanen (Section 2.4,1989). The coding process is briefly described below. First encode the sum $n = \sum_{i=1}^{m} |n_i|$ in a prefix manner. Then, associate with the absolute values of the integers $|n_1|, \cdots, |n_m|$, a binary string. It begins with $|n_1|$ 0's followed by a 1, $|n_2|$ 0's followed by a 1, and so on, until we reach $|n_m|$, for which only $|n_m|$ 0's are added without adding a 1 term. This terminates the string. This string has length $n + m - 1$ and has $m - 1$ 1's. Conversely, any binary string of that length with $m - 1$ 1's defines a set of m non-negative integers. Hence, encoding such sets of integers is equivalent to encoding the binary string associated with them. Define a probability distribution for the set of binary strings of length n such that for each such string x

$$P(x|n) = \frac{m!(n-m)!}{(n+1)!} \tag{1.3.7}$$

where $m = m(x)$ denotes the number of 1's in x. Thus a prefix code can be constructed of length $\log \frac{(n+m)!}{(m-1)!n!}$ for a binary string with length $n+m-1$ and $m-1$ 1's. Finally, using (1.3.7), a prefix code can be constructed with length $\log \frac{(m+1)!}{m_+!(m-m_+)!}$ for the binary string with length m and $m - m_+$ 1's which is used to represent the signs of the integers. In total, we need a prefix code with length (1.3.6) to encode $n_1, \cdots, n_m$.

Because of the relationship amongst the Kraft inequality, the prefix code and the probability distribution, the length function of a prefix code is more important than the code itself. The length function plays an important role in selecting an optimal statistics model.

1.4 Complexity in a Coding System

In Section 1.2 we demonstrated that for any information source $\{A, P\}$ with a finite or countable alphabet and a positive probability function $P(x)$, $-\log P(x)$ is actually established as an optimal length function for a prefix code. Namely, the per-symbol message length achieved by any prefix code cannot improve on that given by $-\log P(x)$ by more than a negligible amount, with probability 1 if n, the length of the message string x, tends to infinity. For such reasons the number

$$I_s(x) = -\log P(x) \tag{1.4.1}$$

is defined to be the Shannon complexity of the string x, relative to an information source $\{A, P\}$. This is the fundamental idea of complexity, although it leaves the crucial part, the information source, unspecified. For each message or event there are necessarily two numbers, the event's probability and its information (the optimal prefix code length), and they are connected by (1.4.1).

An important issue in information theory is the construction of the most suitable information source with which to represent the observed data. As seen above, we can either search for one in terms of probability functions which define random process, or we can look for suitable prefix codes. Such an information theoretic framework is called a coding system, and the associated complexity is fundamental in determining it.

For further study we need to give a precise definition of the coding system. Let A be a finite or countable alphabet. Denote $B = \{0, 1\}$ as the binary alphabet. Then a coding system is defined to be a (decoding) function

$$D : S \to A^* \tag{1.4.2}$$

from a subset S of B^* onto A^*. Usually, the decoding function is not a one-to-one correspondence, which means that each string $x \in A^*$ may have more than one binary string as the codeword to describe it.

A coding system can be constructed from a universal computer, which belongs to the theory of algorithmic complexity, Chaitin (1975), Leung-Yan-Cheong and Cover (1978), Solomonoff(1978) and Zhvonkin and Levin (1970).

A coding system used in this thesis is the one constructed from a family of parametric distribution $\{P(x|\theta), x \in A^*, \theta \in \Theta\}$, where Θ is assumed to be countable. From Section 1.2 for each $\theta \in \Theta$, $-\log P(x|\theta)$ determines a prefix code for A^*. Denote S_θ as the set of all codewords of A^* under a distribution $P(x|\theta)$, and $S_1 = \bigcup_{\theta\in\Theta} S_\theta$. Then it is easy to see that the decoding function $D_1 : S_1 \rightarrow A^*$ is a coding system.

It is not appropriate to simply use the coding system to describe the system generating the observed data, as a redundancy exists in the sense that possibly more than one codeword can describe each data string. We must eliminate this type of redundancy to find a most suitable information source for the observed data generating system.

For the coding system (1.4.2) and an arbitrary $x \in A^*$, denote $D^{-1}(x)$ as the inverse image of x, i.e. $D^{-1}(x)$ is the set of all codewords for x under D. Partition $D^{-1}(x)$ into a class of equivalent sets, where two binary strings u and v are said equivalent if either u is an extension of v or vice versa. If u is an extension of v we say $v < u$ and v is a prefix of u. For a set E of binary strings among which none is a prefix of others, it can be seen $\sum_{u\in E} 2^{-|u|} \leq 1$, where $|u|$ is the length of u. This is the essential property the Kraft inequality characterizes.

Now take the minimal element from each equivalent class, the set of which is denoted as $\bar{D}^{-1}(x)$. It is readily seen that

$$P'(x) = \sum_{u\in\bar{D}^{-1}(x)} 2^{-|u|} \tag{1.4.3}$$

is less than or equal to 1. If $\sum_{x\in A^*} P'(x) = 1$, $P'(x)$ will be a well defined probability distribution on A^*. The associated information source $\{P', A\}$ gives an optimal description for the observed data, which can be seen from the inequality $-\log P'(x) < |u|$ for all $u \in D^{-1}(x)$, i.e. the prefix code associated with $P'(x)$ is the shortest under the coding system D.

If $\sum_{x\in A^*} P'(x) < 1$, we define a probability distribution $P(x)$ and accordingly an information source $\{A, P\}$, by the recursive normalization

$$P(x^0) = 1, P(x^{n+1}) = \frac{P(x^n)P'(x^{n+1})}{\sum_{z\in A} P'(x^n, z)} \tag{1.4.4}$$

where x^n, z denotes the string of length $n+1$ formed by concatenating x^n with the symbol z. By this way the prefix code length $-\log P(x)$ can be further shortened from $-\log P'(x)$.

The complexity of x, relative to the coding system D, as defined by Rissanen (1989), is

$$I(x|D) = -\log P(x). \tag{1.4.5}$$

From (1.4.5) and (1.4.3) it is easily shown that the prefix code length $-\log P(x)$ for x is shorter than the length of any code of x in the coding system. Therefore by describing x in terms of $-\log P(x)$ we remove the redundancy of the coding system and obtain a shortest description.

1.5 Two-part Codes

In this section, we give the description of the data generated from an unknown probabilistic model which belongs to an assumed class. The description procedure is the so-called two-step encoding process, Rissanen (1989).

Let the assumed model class be denoted by

$$M_1 = M_1(k) = \{p(x|\theta), \pi(\theta)\} \tag{1.5.1}$$

where x takes values in a measurable space $\mathcal{X}$ and $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ denotes a k-component parameter vector ranging over a closed subregion Ω_k of the k-dimensional Euclidean space $\mathcal{R}^k$. Here k is the dimension of the parameter θ and we treat it as a parameter. Suppose $p(\cdot|\theta)$ is an almost surely positive density function with respect to a known complete, σ-finite dominating measure $\nu(\cdot)$, so that the usual discrete and continuous distributions are included. Here the $\pi(\theta)$ is a probability density of θ on

Ω_k, i.e. $\int_{\Omega_k} \pi(\theta)d\theta = 1$. It is classically regarded as a prior density but later we will give an information theoretic interpretation.

Now the description of the data depends on the selection of the model from M_1. It is natural to conduct the description in two steps. First use a prefix code to describe a model of M_1 from which the data is assumed to be generated. Next, encode the data into another prefix code using the assumed model, and then take the concatenation of the two prefix codes as the description of the data.

However we have to overcome some difficulties before we can construct any prefix code for the data. The prefix codes are built on the data or the message in which each term ranges over a finite or countable alphabet A. When each data item ranges over an uncountable set, as in the case of continuous distribution, there will be no finite-length prefix code for the data. Such a problem also exists in describing the models of M_1 if its size is uncountable.

To overcome these difficulties a quantization process should be employed before conducting the two-step encoding process. A class of discrete distributions, M_1', whose size is countable can be constructed based on the quantization to approximately represent the model class M_1. Then the two-step encoding applied on M_1' gives us a two-part code for the data x and this two-part code can be optimized by the optimal parameters and quantization. We regard the resultant optimal two-part code as the description of x relative to M_1.

The quantization for the data x is natural since x is usually observed to a prescribed precision. However the quantization for the parameters θ must be optimized.

Denote $[x]$ as the quantization region that contains x; and $d = \nu([x])$ as the precision of x. Suppose the data x is observed to the precision d, then in each $[x]$ only one x can be observed and there is no confusion using $[x]$ to represent it. We write $[\mathcal{X}]$ as the whole set of $[x]$. Suppose, in addition, we truncate the parameters $\theta = (\theta_1, \cdots, \theta_k)$ to $\bar{\theta} = (\bar{\theta}_1, \cdots, \bar{\theta}_k)$ to some precision $\delta = (\delta_1, \cdots, \delta_k)$, and write the whole set of $\bar{\theta}$ as $\bar{\Omega}_k$. With this notation, it is easy to see that

$$M_1' = \{P([x]|\bar{\theta}), \Pi(\bar{\theta})\} \tag{1.5.2}$$

is a class of discrete distributions, the size of which is countable. Here $P([x]|\bar{\theta}) = \int_{[x]} p(x|\bar{\theta})\nu(dx) \approx p(x|\bar{\theta})\nu([x])$ and $\Pi(\bar{\theta}) \approx \pi(\bar{\theta})\prod_{i=1}^{k}\delta_i$ if the quantization is sufficiently fine. Now we construct a prefix code $C(\bar{\theta})$ for each $\bar{\theta} \in \bar{\Omega}_k$ by using the results in Section 1.2, the code length of which is given by $L_1(\bar{\theta}) = -\log \Pi(\bar{\theta})$. Similarly, we can construct a prefix code $C(x|\bar{\theta})$, the length of which is $L_1(x|\bar{\theta}) = -\log P([x]|\bar{\theta})$, for the observed data x if we employ $P([x]|\bar{\theta})$ as the generating distribution. The resultant two-part code

$$C(x, \bar{\theta}) = C(\bar{\theta})C(x|\bar{\theta}) \tag{1.5.3}$$

can then be seen as a natural description of x relative to $P([x]|\bar{\theta})$ and its total code length is given by

$$\begin{aligned} L_1(x, \bar{\theta}) &= L_1(\bar{\theta})) + L_1(x|\bar{\theta}) \\ &= -\log P([x]|\bar{\theta}) - \log \Pi(\bar{\theta}) \\ &= -\log p(x|\bar{\theta}) - \log \pi(\bar{\theta}) - \sum_{i=1}^{k} \log \delta_i - \log \nu([x]) + o(1) \end{aligned} \tag{1.5.4}$$

if the quantization is sufficiently fine and $p(x|\theta)$ and $\pi(\theta)$ satisfy some smooth conditions. Usually the precision d for x is already implied in the observations so $\nu([x])$ is a constant. As our objective is to find a measure for model selection, we choose only the dominating terms in (1.5.4) and define

$$L_1(x, \bar{\theta}) = -\log p(x|\bar{\theta}) - \log \pi(\bar{\theta}) - \sum_{i=1}^{k} \log \delta_i \tag{1.5.5}$$

as the two-part code length for x relative to $p(x|\bar{\theta})$.

Among all the possible code lengths $L_1(x, \bar{\theta})$, where $\bar{\theta} \in \bar{\Omega}_k$, it is natural to choose the one with the smallest length as the most suitable description of the data x, if we only consider two-part encoding process; namely, we choose a $\hat{\bar{\theta}}$ and its corresponding information source $\{p(x|\hat{\bar{\theta}}), [\mathcal{X}]\}$ as if the data x is generated from it so that it achieves the minimum two-part code length

$$\min_{\bar{\theta}} \left\{ -\log p(x|\bar{\theta}) - \log \pi(\bar{\theta}) - \sum_{i=1}^{k} \log \delta_i \right\}. \tag{1.5.6}$$

The rationale for this choice of $\bar{\theta}$ lies on the results of Section 1.2 where we have already demonstrated that $-\log Q(x)$ gives the optimal code length for x if Q is the distribution generating x.

Suppose the probability density functions $p(x|\theta)$ and $\pi(\theta)$ are smooth enough so that the minimization of $L_1(x,\theta)$ is achieved at one finite point $\hat{\theta}$. Now $L_1(x,\theta)$ can be expanded in Taylor's series around $\hat{\theta}$ and we can get information about the optimizing precisions. Note that, subject to the smooth conditions, $\hat{\bar{\theta}}$ which achieves (1.5.6) is close to $\hat{\theta}$ within the truncation precision δ. Therefore by Taylor's expansion

$$\begin{aligned} L_1(x,\hat{\bar{\theta}}) &= L_1(x,\hat{\theta}) + \frac{1}{2}(\hat{\bar{\theta}}-\hat{\theta})\Sigma(\hat{\bar{\theta}}-\hat{\theta})^T \\ &\leq -\log p(x|\hat{\theta}) - \log \pi(\hat{\theta}) + \frac{1}{2}\delta\Sigma\delta^T - \sum_{i=1}^{k}\log\delta_i \end{aligned} \tag{1.5.7}$$

where Σ denotes the matrix of the double derivatives of the function $L_1(x,\theta)$ with respect to θ evaluated at some point near $\hat{\theta}$. By taking derivative with respect to δ of the right hand side of (1.5.7), a minimax upper bound for $L_1(x,\hat{\bar{\theta}})$ is

$$\begin{aligned} L_1(x,\hat{\theta},\hat{\delta}) &= -\log p(x|\hat{\theta}) - \log \pi(\hat{\theta}) + \frac{1}{2}\hat{\delta}\Sigma\hat{\delta}^T - \sum_{i=1}^{k}\log\hat{\delta}_i \\ &= -\log p(x|\hat{\theta}) - \log \pi(\hat{\theta}) + \frac{\ln 2}{2}k - \sum_{i=1}^{k}\log\hat{\delta}_i \end{aligned} \tag{1.5.8}$$

where $\hat{\delta}$ is the solution of the equation

$$(\delta_1,\cdots,\delta_k)\Sigma = (\delta_1^{-1},\cdots,\delta_k^{-1})\ln 2. \tag{1.5.9}$$

The estimated precision $\hat{\delta}$ is optimal, and by (1.5.7) $L_1(x,\hat{\theta},\hat{\delta})$ gives us an optimal worst case two-part code length of the data x. Since (1.5.8) does not involve the precision d and the truncation precision δ for θ is optimized, we also regard it as the optimal two-part code length for x relative to M_1.

In statistical model selection we are often interested in selecting an optimal number of parameters for the assumed model. From (1.5.8) we can derive, under some conditions, an asymptotic approximation which is quite useful for solving this dimension selection problem. Suppose that $-\log p(x|\theta)$ grows proportionally to the

number of observations n, which is normally satisfied in most situations, the elements of $S = \Sigma/n$ are of the order of 1 regardless of n. Then from (1.5.9) $\hat{\delta}_i = c_i(n)/\sqrt{n}$, where $c_i(n)$ is finitely bounded, and the expression (1.5.8) simplifies to the form

$$MDL_1(k) = -\log\left(p(x|\hat{\theta})\pi(\hat{\theta})\right) + \frac{k}{2}\ln n + O(k). \tag{1.5.10}$$

For large number of observations, (1.5.10) with the last term $O(k)$ removed could serve as a criterion to select the optimal dimension of the model, which is called the MDL (minimum description length) principle. See Section 1.8 for more discussion.

In addition to the usual Bayesian interpretation, $\pi(\theta)$ can also be described as a measure for the complexity of the model and therefore an order of preference: we would prefer a model which is as simple as possible subject to the requirement that it provides an efficient description for the data. Technical advantages of using $\pi(\theta)$ can be found in the next section where we obtain a closed form of the stochastic complexity from the coding system which was introduced in the previous section.

If the data x comes from a distribution without a prior distribution we can use the universal prior of integers to describe the distribution and then give the two-part code length. This was done by Rissanen (1983), and described briefly as follows.

Suppose the model class for the data x is

$$M_2 = M_2(x) = \{p(x|\theta)\} \tag{1.5.11}$$

where $x \in \mathcal{X}$, $\theta \in \Omega_k$ and $p(x|\theta)$ is a density function all defined the same as for M_1. Assume moreover that the usual smooth conditions for $p(x|\theta)$ hold so that the maximum likelihood estimate $\hat{\theta}$ exists and the Taylor's expansion of $-\log p(x|\theta)$ around $\hat{\theta}$ is available.

To describe the data x, we need first to describe θ in a prefix manner and then describe x with the employed density $p(x|\theta)$. The resulting two-part code length is $L_2(x,\theta) = L_2(x|\theta) + L_2(\theta)$. For a fixed θ, $L_2(x|\theta)$ is equivalent to $-\log p(x|\theta)$ except for a constant. This constant is completely determined by the natural precision of x, and not important for the code length. With the same argument as before, we still need a quantization for Ω_k, otherwise no finite uniquely decodable code exists.

Usually we want the first term in $L_2(x,\theta)$ to be dominant. We also want the minimizing values for each set of parameters to be close to the maximum likelihood estimates. The problem thus is to decide on the precision to be used for the maximum likelihood estimates of θ. Clearly, if we use a coarse precision, the second term $L_2(\theta)$ in $L_2(x,\theta)$ will be small, but the first term will grow from its minimum, since we are generally no longer using the correct maximum likelihood estimates due to the truncation.

Keeping these in mind, we partition Ω_k into a set of identical k-dimensional parallelepipeds and truncate every θ in Ω_k to the center $\bar{\theta}$ of the parallelepiped in which it falls. The truncation precision δ for the parameter θ is determined by

$$\delta M(\hat{\theta})\delta^T = \gamma \tag{1.5.12}$$

which comes from the Taylor's expansion

$$-\log p(x|\bar{\hat{\theta}}) \approx -\log p(x|\hat{\theta})) + \frac{1}{2}(\bar{\hat{\theta}} - \hat{\theta})M(\hat{\theta})(\bar{\hat{\theta}} - \hat{\theta})^T \tag{1.5.13}$$

in which we wish to control the second term by a temporarily prescribed γ. Here $M(\theta)$ is the matrix of the double derivatives of the function $-\log p(x|\theta))$, and $\bar{\hat{\theta}}$ is the center of the parallelepiped containing $\hat{\theta}$.

Now we start from the parallelepiped which is in volume the largest inscribed rectangle of the ellipsoid $(\theta - \hat{\theta})M(\hat{\theta})(\theta - \hat{\theta})^T \leq \gamma$. Its the volume is $V(\gamma) = (4\gamma/k)^{k/2}\sqrt{\det M(\hat{\theta})}$. Shift this parallelepiped respectively along each of its k sides by the corresponding distance $2\sqrt{\gamma/(k\lambda_i)}$, where λ_i denotes the i-th eigenvalue of $M(\hat{\theta}))$, then proceed with the same shifting for each of the k new parallelepiped and continue the operation until Ω_k is covered by these parallelepipeds. The resulting set is denoted by $\bar{\Omega}_k = \{\bar{\theta}, \theta \in \Omega_k\}$. Next we order these parallelepipeds by assigning an integer index to each of them. This can be done by using the natural distance $\bar{\theta}M(\bar{\theta})\bar{\theta}^T$ and enumerating $\bar{\Omega}_k$ in a right-handed system. As a consequence of this enumeration the index $n(\bar{\theta})$ is given approximately by the ratio of the volume enclosed by the ellipsoid $\{y : yM(\hat{\theta})y^T \leq \bar{\theta}M(\hat{\theta})\bar{\theta}^T\}$ to the volume $V(\gamma)$, i.e. $n(\bar{\theta}) = C_k(k\bar{\theta}M(\hat{\theta})\bar{\theta}^T/(4\gamma))^{k/2}$,

where C_k is defined as the volume of the k-dimensional unit ball, which equals to

$$C_k = \begin{cases} (2\pi)^{k/2} / [(k/2)! 2^{k/2}] & k \text{ even}, \\ \pi^{(k-1)/2} 2^{k+1} ((k+1)/2)! / (k+1)! & k \text{ odd}. \end{cases} \tag{1.5.14}$$

Applying the universal prior to $\{n(\bar{\theta}), \bar{\theta} \in \bar{\Omega}_k\}$, we can find a prefix code for $\bar{\Omega}_k$ the length of which equals to $L^*(n(\bar{\theta})) = \log^* n(\bar{\theta}) + \log 2.865$.

With the quantization of Ω_k and the universal prior for $n(\bar{\theta})$, the two-part code length for x relative to M_2 is given by

$$\begin{aligned} L_2(x,\theta) &= -\log p(x|\bar{\hat{\theta}}) + L^*(n(\bar{\hat{\theta}})) \\ &\leq -\log p(x|\hat{\theta}) + \frac{1}{2}\gamma + \log^* n(\bar{\hat{\theta}}) + \log 2.865. \end{aligned} \tag{1.5.15}$$

We may ask for the value of γ which minimizes the quantity

$$\frac{1}{2}\gamma + \log^* n(\bar{\hat{\theta}}). \tag{1.5.16}$$

If we approximate $\log^*$ by log, the optimum value for γ is $k \log e$. Substituting the optimum γ into the right hand side of (1.5.15) the minimum upper bound for $L_2(x,\theta)$ is

$$\begin{aligned} &-\log p(x|\hat{\theta}) + \log C_k + \frac{k}{2} \log \theta M(\hat{\theta}) \hat{\theta}^T + O(k) \\ &= -\log p(x|\hat{\theta}) + \frac{k}{2} \log \frac{n}{k} + k \log ||\hat{\theta}||_{I(\hat{\theta})} + O(k) \end{aligned} \tag{1.5.17}$$

where $||\hat{\theta}||_{I(\hat{\theta})} = \sqrt{\hat{\theta} M(\hat{\theta}) \hat{\theta}^T / n}$ and $I(\hat{\theta}) = M(\hat{\theta})/n$. The expression (1.5.17) gives us an optimal worst case two-part code length and we regard it as the minimum two-part code length of x relative to M_2.

Assuming that $-\log p(x|\theta)$ grows proportionally to the number of the observations n, the minimum two-part code length simplifies to the form

$$MDL_2(k)) = -\log p(x|\hat{\theta}) + \frac{k}{2} \log n + O(k). \tag{1.5.18}$$

Therefore for large number of observations, the right hand side of (1.5.18) with the last term removed can be used as a criterion for dimension selection.

In addition to the two-part code length function, discussed here for parametric models both with or without a prior distribution, it is also possible to construct a two-part code length function for nonparametric model classes. For details see Rissanen et al. (1992), Yu and Speed (1992), Speed and Yu (1992) and Hall and Hannan (1988). We will also discuss this case in Chapter 5.

1.6 Stochastic Complexity

From the two-part encoding procedure discussed in the last section, a coding system can be obtained relative to model class M_1 or M_2 which is defined as

$$D : S \to [\mathcal{X}].$$

Here S is the set of codewords $C(x, \bar{\theta}) = C(\bar{\theta})C(x|\bar{\theta})$, where $C(\bar{\theta})$ is the prefix code for the truncated parameter $\bar{\vartheta}$, and $C(x|\bar{\theta})$ is the prefix code for x under the employed model $p(x|\theta)$. Similar to that in Section 1.4, we can construct an information source by which the data is described with the shortest code length relative to M_1 or M_2. One may ask the relationship between this shortest code length and the optimal two-part code lengths derived in Section 1.5.

Obviously the length of the two-part code $C(x, \bar{\theta})$ is longer, irrespective of the value of θ, since $C(x, \bar{\theta})$ actually gives us more than we want. We initially set out to encode the data x, and ended up encoding both the data and some parameters. On the other hand, the optimal two-part code length relative to M_1 (or M_2) is the one among those of $C(x, \bar{\theta})$ which is the closest to the shortest code length. In fact, as it will be shown later, they are equal in an asymptotic sense under some smoothness conditions for the model distribution.

First we derive the shortest code length relative to M_1 following the procedure in Section 1.4. Since the two-part code $C(x, \bar{\theta})$ is prefix, we may substitute the code length (1.5.5) into (1.4.3) with the result

$$p'(x) = \sum_{\bar{\theta}} 2^{-L_1(x,\bar{\theta})} = \sum_{\bar{\theta}} p(x|\bar{\theta})\pi(\bar{\theta}) \prod_{i=1}^{k} \delta_i \tag{1.6.1}$$

where the summation is taken over all the truncated parameter values $\tilde{\theta}$. $p'(x)$ is already a probability density function so there is no need for the recursive normalization (1.4.4). Now letting $\delta \to 0$ the sum in (1.6.1) goes over to the integral

$$p(x) = \int_\Omega p(x|\theta)\pi(\theta)d\theta. \tag{1.6.2}$$

From (1.4.5), $-\log p(x)$ is the shortest code length relative to the coding system D. Hence we define the stochastic complexity of the data x, relative to the model class M_1, as

$$I(x|M_1) = -\log p(x) = -\log \int_\Omega p(x|\theta)\pi(\theta)d\theta. \tag{1.6.3}$$

The fact that the code length $-\log p(x)$ was obtained by the removal of a redundancy in the coding system, which is defined by the model class, lends it a natural sense of minimality, which is certainly difficult to achieve otherwise. That (1.6.3) is smaller than the optimal two-part code length is easy to see: the sum (1.6.1) is clearly larger than any of its terms, including the maximum.

Suppose $p(x|\theta)\pi(\theta)$ is smooth enough so its logarithm admits Taylor's expansion about $\hat{\theta}$ maximizing $p(x|\theta)\pi(\theta)$, i.e.

$$\log p(x|\theta)\pi(\theta) = \log p(x|\hat{\theta})\pi(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})\hat{\Sigma}(\theta - \hat{\theta})^T$$

where $\hat{\Sigma}$ is the Hessian matrix of the double derivatives of $-\log p(x|\theta)\pi(\theta)$ evaluated at some point near $\hat{\theta}$. Then $p(x|\theta)\pi(\theta) = p(x|\hat{\theta})\pi(\hat{\theta})2^{-\frac{1}{2}(\theta-\hat{\theta})\hat{\Sigma}(\theta-\hat{\theta})^T}$ and

$$p(x) = p(x|\hat{\theta})\pi(\hat{\theta}) \int 2^{-\frac{1}{2}(\theta-\hat{\theta})\hat{\Sigma}(\theta-\hat{\theta})^T} d\theta = p(x|\hat{\theta})\pi(\hat{\theta})|\hat{\Sigma}|^{\frac{1}{2}} O^k(1).$$

So

$$\begin{aligned} I(x|M_1) &= -\log p(x) = -\log p(x|\hat{\theta})\pi(\hat{\theta}) + \frac{1}{2}\log|\hat{\Sigma}| + O(k) \\ &= -\log p(x) = -\log p(x|\hat{\theta})\pi(\hat{\theta}) + \frac{k}{2}\log n + O(k) \end{aligned} \tag{1.6.4}$$

if $\hat{\Sigma}$ is of order n. The asymptotic equivalence between the stochastic complexity (1.6.3) and the optimal two-part code length (1.5.10) under some smoothness conditions is clearly visible now.

From this asymptotic equivalence we conclude that even though the two-part encoding procedure for the data x is ad hoc in the sense that it involves redundancy in the description, it reduces the redundancy to a minimal, negligible amount when the procedure is optimized. The use of the stochastic complexity and the optimal two-part code length for model selection is further discussed in Section 1.8.

The stochastic complexity relative to the model class M_2 was derived recently by Rissanen (1994a). By taking into account the Fisher information and removing an inherent redundancy in the two-part codes a sharper code length is given as the stochastic complexity. In Section 1.5 we described the two-part code for M_2 which is computed by Rissanen (1983). There the data are encoded with the maximum likelihood model and preceded by the encoded parameters $\bar{\hat{\theta}}$ truncated to a precision δ. Here this procedure is refined in two ways. First, the truncation is made to be dependent on equivalent classes $R(\bar{\hat{\theta}})$ which are determined by the Fisher information. Then with

$$P_{\bar{\hat{\theta}}} = \int_{\hat{\theta}(x) \in R(\bar{\hat{\theta}})} p(x|\bar{\hat{\theta}}) \nu(dx) \tag{1.6.5}$$

an inherent redundancy in the earlier procedure is removed, and the total nonredundant two-part code length is given by

$$L(x, \delta) = -\log \frac{p(x|\bar{\hat{\theta}})}{P_{\bar{\hat{\theta}}}} + L(\bar{\hat{\theta}}). \tag{1.6.6}$$

There is no longer any optimal precision, and the shortest worst case code length results from the infinite precision $\delta = 0$. Since it is nonredundant, and can be approximated by a two-part code with error as small as possible, we regard this shortest code length as the stochastic complexity relative to M_2. Under the main condition that the maximum likelihood estimates satisfy the central limit theorem, an asymptotic expression for this stochastic complexity is given by

$$I(x|M_2) = -\log p(x|\hat{\theta}) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int_\Omega \sqrt{I(\theta)} d\theta + o(1) \tag{1.6.7}$$

where $I(\theta)$ is the Fisher information matrix

$$I(\theta) = \left\{ -E \frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j} \right\}.$$

See Rissanen (1993) for further details.

1.7 Predictive Coding

As an alternative to the two-step encoding process, there is another encoding process, called predictive coding, which describes the data generated from an unknown probabilistic model. In this process, the encoder does not need to provide a prefix code for the model (usually it turns out to describe the parameters that characterize the model). Instead he estimates the parameters characterizing the model from the available data according to an optimal procedure known to the decoder. Then writes a prefix code for the next observation based on this model fitted from the previous data. Each time the encoder obtains a new observation he updates the estimate of the parameters and encode the next observation with the latest fitted model. The predictive coding process removes the redundancy of the coding system in a way that is quite different from that of the stochastic complexity. The resulting predictive code length for the observed data string is called the predictive stochastic complexity.

The predictive stochastic complexity and, accordingly, the predictive minimum description length principle, was proposed and studied by Rissanen (1986a). However, the predictive process was also discovered by Dawid (1984,1991b), who proposed it as a prequential method for probabilistic forecasting. There is also a closely related technique by Hjorth (1982), called forward validation, whose main objective is to reduce the bias in the estimates of the variance of parameter estimators.

A brief overview of the derivation of the predictive complexity is given below. Without loss of generality, we consider only the model class M_2 introduced in Section 1.5. Rewrite M_2 as

$$M_2 = M_2(k) = \{p_{k,\theta}(x), x \in \mathcal{X}, \theta \in \Omega_k\}, \tag{1.7.1}$$

and denote $x = x_1, \cdots, x_n = x^n$ as a sample of observations generated from an unknown density function belonging to M_2. To be general, x is assumed to be a

random process satisfying the compatibility condition

$$\int p_{k,\theta}(x_1, \cdots, x_t, z)\nu(dz) = p_{k,\theta}(x_1, \cdots, x_t). \tag{1.7.2}$$

We now proceed to describe x using a predictive coding process. Predictive coding means that we want to find the conditional density for the next observation x_{t+1} (regarded as a random variable) based upon the previous observations $x_1 \cdots, x_t$,

$$p_{k,\hat{\theta}(t)}(x_{t+1}|x_1, \cdots, x_t) \tag{1.7.3}$$

where $\hat{\theta}(t)$ is estimated from $x_1, \cdots, x_t$ using a procedure known to the decoder. With this conditional density the minimum code length needed to encode x_{t+1} in a prefix manner is $-\log p_{k,\hat{\theta}(t)}(x_{t+1}|x_1, \cdots, x_t)$. The total code length is

$$L(x|k) = -\sum_{t=0}^{n-1} \log p_{k,\hat{\theta}(t)}(x_{t+1}|x_1, \cdots, x_t). \tag{1.7.4}$$

This may be minimized with respect to k to give the estimate $\hat{k}(n) = \hat{k}(x^n)$.

The code length (1.7.4) does not provide a complete description of x since the information about the dimension k of the parameter θ is unknown to the decoder. Therefore another prefix code for k is required, the optimal length of which is $L^*(k) = \log^* k + \log c^*$, the one defining the universal prior of the integers in Section 1.3. We call the corresponding minimum code length

$$I_{sp}(x|M_2) = \min_k \{L(x|k) + L^*(k)\} \tag{1.7.5}$$

the semi predictive stochastic complexity. The word "semi" suggests that the optimal dimension $\hat{k}(n)$ is not determined the predictive way. Still, using $\hat{k}(n)$ to denote the optimal k is to emphasize that the main factor of determining the dimension of the parameters in (1.7.5) is the first term, and in almost all the cases of interest the minimizations of (1.7.4) and (1.7.5) produce exactly the same dimension of the parameters.

Modifying the above procedure of describing the data sequence a purely predictive stochastic complexity can be defined as

$$I_p(x|M_2) = -\sum_{t=0}^{n-1} \log p_{\hat{k}(t),\hat{\theta}(t)}(x_{t+1}|x_1, \cdots, x_t). \tag{1.7.6}$$

At each time t the k in (1.7.4) is now replaced by the optimal $\hat{k}(t)$, which minimizes $L(x^t|k)$, so that an optimal conditional density $p_{\hat{k}(t),\hat{\theta}(t)}(x_{t+1}|x_1,\cdots,x_t)$ is obtained to encode the next observation x_{t+1} with the optimal code length $-\log p_{\hat{k}(t),\hat{\theta}(t)}(x_{t+1}|\, x_1, \cdots, x_t)$. Therefore, (1.7.6) completely represents the description of the data x relative to M_2, and we do not need the prefix code for the dimension of the parameters because the algorithm of determining it is already known by the decoder.

The selection of the optimal estimate $\hat{\theta}(t)$ in (1.7.4) for each k proceeds as follows. One might think of choosing $\hat{\theta}$ so that the code length for x_{t+1}, $-\log p_{k,\theta}(x_{t+1}|x_1, \cdots, x_t)$, is minimized. But such $\hat{\theta}$ would be a function of x_{t+1} which would make the decoding impossible. To avoid this, we apply the essential idea of inductive inference: In the light of past observations the best single value of the parameter for encoding the "next" observations, x_{i+1}, $i = 0, 1, \cdots, t-1$ is the value that minimizes the sum $-\sum_{i=0}^{t-1} \log p_{k,\theta}(x_{i+1}|x_1, \cdots, x_i)$, i.e. the maximum likelihood estimate $\hat{\theta}$ (Rissanen,1986a). Such a selection of $\hat{\theta}$ is based on the hope that the prediction distribution (1.7.3) for the new observation x_{t+1} is like it was in the past.

To carry out the computation of the predictive complexity for a data sequence there are still several points needed to be clarified, i.e. the order of the data sequence ((1.7.4),(1.7.5) and (1.7.6) are affected by the order of the data, especially the order of the first few data points) and the initial estimate $\hat{\theta}(0)$. For details of these issues, refer to Rissanen (1986a and Chapter 5 of 1989) and Section 4.4 of this thesis.

Understanding the asymptotic behavior of stochastic complexity is helpful in studying the optimal properties of the model selection by stochastic complexity. Results for the asymptotic behavior for several model classes are available in the literature. See Rissanen (1986a, 1987, 1989) for the parametric density class; Rissanen (1986b) for the class of Markov chains; Rissanen (1986c) and Speed and Yu (1993) for the Gaussian regression problem; Rissanen et al. (1992), Hall and Hannan (1988) and Yu and Speed (1992) for the nonparametric density class; Rissanen (Chapter 6, 1989), Hannan et al. (1989), Hemerly and Davis (1989), Gerencsér (1989,1992) and Gerencsér and Rissanen (1992) for time series. Also a related work is Barron and

Cover (1991) which applies the algorithmic complexity to density estimation.

Here we give some of the important results concerning the stochastic complexity relative to a parametric density class.

Theorem 1.7.1 *(Rissanen, 1986a). Let for each k the parameters θ range over a compact subset Ω_k with nonempty interior of the k-dimensional Euclidean space. We assume that there exist estimates $\hat{\theta}(x^n)$ satisfying the central limit theorem such that the tail probabilities are uniformly summable as follows*

$$P_\theta\left(\sqrt{n}||\hat{\theta}(x^n) - \theta|| \geq \log n\right) \leq \delta(n) \quad \textit{for all } \theta \textit{ and } \textstyle\sum_n \delta(n) < \infty. \tag{1.7.7}$$

where $||\theta||$ denotes a norm. If g is any density defined on the observations, satisfying the compatibility conditions for a random process, then for all k and all $\theta \in \Omega_k$, except in a set of Lebesgue measure zero,

$$\liminf_{n\to\infty} \frac{E_{k,\theta} \log [p_{k,\theta}(x^n)/g(x^n)]}{(k/2)\log n} \geq 1. \tag{1.7.8}$$

The mean is taken relative to the distribution defined by $p_{k,\theta}$.

This theorem states that for all k, all positive number ε, and for all points $\theta \in \Omega_k$, except in a null set,

$$E_{k,\theta} \log \frac{p_{k,\theta}(x^n)}{g(x^n)} \geq (\frac{1}{2} - \varepsilon) k \log n. \tag{1.7.9}$$

This is a generalization of Shannon's famous coding theorem, in that the average prefix code length $E_{k,\theta} \log g(x^n)$ is not only greater than or equal to the entropy but exceeds it by a positive number, which represents the amount of uncertainty in the class of models. From (1.7.9) it follows that the minimum two-part code length in Section 1.5 and the stochastic complexity in Section 1.6 both reach asymptotically the minimum bound, provided that the model densities satisfy certain mild smoothness conditions. This gives a rational basis for using stochastic complexity as a model assessment measure.

It has also been proved that the minimum bound in (1.7.9) can be achieved for the semi predictive complexity if the data points are independent. This result is stated below.

Theorem 1.7.2 *(Rissanen, 1986a). Let the family of densities satisfy the conditions for independence for each k and $\theta \in \Omega_k$, namely, $p_{k,\theta}(x) = \prod_{t=1}^{n} p_{k,\theta}(x_t)$, and let $_{k,\theta}(x_t)$ be three times continuously differentiable with respect to θ in the interior of a compact set Ω_k. Further, let the central limit theorem hold for some estimates $\hat{\theta}(x^n)$ of θ in the interior points such that the four first moments of $\sqrt{n}(\hat{\theta}(x^n) - \theta)$ converge. Then $I_{sp}(x^n|M_2)$, defined by the equation (1.7.5), is optimal in that for all k and all θ in Ω_k,*

$$I_{sp}(x^n|M_2) \leq -E_{k,\theta} \log p_{k,\theta}(x_n) + \frac{k}{2} \log n + o(\log n). \tag{1.7.10}$$

If the model class is nonparametric and the density g in Theorem 1.7.1 is restricted to be histogram type, then the minimax bound of $E_p \log \frac{p(x^n)}{g(x^n)}$ has been shown to be of order $n^{1/3}$ assuming $x_1, \cdots, x_n$ is a simple random sample. This bound can also be achieved both in expectation and almost surely by histogram densities induced by the predictive stochastic complexity (see Yu and Speed (1992), Rissanen et al. (1992) and Barron and Cover (1991) for detail).

1.8 Principle of Minimum Description Length

In traditional statistical inference if a probability density $p(\cdot|\theta)$ for the observed data string x is given, where θ, with its dimension fixed, belongs to a parametric space Ω, then one of the most important methods to estimate the unknown parameter is the maximum likelihood principle. The traditional measure for goodness of the estimators is their variance, or some other related utility function. An unbiased estimator is called efficient if its covariance achieves the lower bound set by the Cramer-Rao inequality. Numerous results of the consistency and asymptotic efficiency of the maximum likelihood estimate can be found in the literature, e.g. Lehmann (1986a). If we consider a prior distribution for the parameter, the parameter estimation may be carried out in many cases by the maximum posterior or the ML-II technique (Good (1983) and Berger (1985)).

However, if the dimension of the parameter θ is unknown and is to be estimated, as in a regression model, time series model and the unsupervised classification problem, the above traditional methods for parameter estimation do not work. Instead, the traditional hypothesis testing technique is used, like the likelihood ratio test. But in many cases the hypothesis testing procedure has some unsatisfactory features: the selection of the level of significance, which should depend on the amount of data, is subjective; most powerful test usually does not exist, etc. Needless to say, there are remedies for estimating the dimension of the parameter. Two widespread methods are Akaike's AIC, Akaike (1970, 1974a, 1974b,1977) and the cross-validation technique, Geisser and Eddy (1979) Stone (1974,1977a,1977b). In this section, however, we introduce the competent minimum description length (MDL) principle, Rissanen (1986a,1987,1989).

The MDL principle contains a three-level hierarchy of modeling problems and a redevelopment of estimation theory. On the lowest level, it assumes the model class generating the data string to be a set of probabilistic distributions with the fixed number of parameters. The task is to find good or even optimal estimates of the unknown parameters. The choice of the parameter estimate is the one that minimizes the code length relative to the model class. From the process of deriving (1.5.10), (1.5.18) and (1.7.4) we note that the proposed parameter estimates are exactly or asymptotically the same as the maximum likelihood estimates or the maximum posterior estimates. We regard the resulting shortest code length as the stochastic complexity even though we have already used the concept in Section 1.6. These, however, are shown to be equivalent to each other in the asymptotic sense.

On the next level in the hierarchy, the model class assumed to generate the data string is generalized to be a family of model classes, each of which has its own dimension for the parameters. At this stage, the traditional estimating methods do not work while the traditional hypothesis testing procedure lacks the sufficient capability. However, by looking for the shortest description of the data under the current model family an optimal estimate of the dimension of the parameters, as well as their

estimates, can be obtained. Notice that while there may exist a model class under which the stochastic complexity of the observed data is very small, the code length required to describe the model class itself could be very large. An example of an extreme case is when we put probability 1 on the observed x and 0 elsewhere, we need 0 length code to describe x under such a model. But the complexity of the model itself would be so large that it can not be specified unless we know the true distribution of x. The stochastic complexity and the predictive stochastic complexity described in the previous sections (like (1.5.10), (1.5.18), (1.6.7) and (1.7.5)) give a criterion under which the optimal dimension of the parameter can be obtained by balancing the complexity of the data under the employed model and the complexity of the employed model class in the best possible manner. After the optimal dimension is found, the stochastic complexity relative to the current model family can also be calculated as the shortest two-part code length under the specified model family, namely, one part for the stochastic complexity relative to a model class and the other for the complexity of the employed model class.

Using the MDL principle for the dimension estimation has been found to be very successful in many statistical problems, such as the regression modeling (where the MDL principle is known as the predictive least square and the predictive least quasi-deviance), times series, classification and density estimation. The consistency of the dimension estimate has been proved for several modeling problems (see Rissanen (1986c, 1986d, 1989), Hannan et al. (1989), Hemerly and Davis (1989) and Gerencsér (1989,1992)).

Finally, an attempt is made to find other possible model families so that a better or even the best model for generating the observed data string may be found. While such problem goes completely beyond the reach of traditional statistics, it has important use in practice. For example, in regression analysis there are several useful estimates for the regression model: the simple linear regression estimator, the polynomial spline regression estimator, the projection pursuit regression estimator and other nonlinear regression estimators. Not only do we want to find an optimal regression model

estimate under each regression method, but we also want to compare these estimates to obtain the best model estimate. The stochastic complexity or the shortest abstract code length gives a global measure to compare these different estimates. We can begin by finding the stochastic complexity of the data relative to each model family. Then calculate the abstract two-part code length for the data relative to a set of model families, provided that an algorithm is available to describe this set of model families. In practice it is not difficult to find the shortest code length to describe the model families because we usually study a preselected set of families. Theoretically or conceptually, however, finding a shortest code length for the description of the model families is extremely difficult and it depends critically on the particular formalization of the ground language. For the detailed discussion, refer to Section 3.6 of Rissanen (1989).

The complete statistical estimation or modeling problem, which consists of the above three levels, is handled by the minimum description length principle in a uniform manner. Beginning from the highest level, we search for the model family which results in the minimum abstract code length relative to a set of preselected model families for the data. We call it $\mathcal{M}^*$, the best family we know. Next we seek the best model class to minimize the two-part code length of the data relative to $\mathcal{M}^*$, which is specified in terms of the best dimension of the parameters k^*. The best model class is denoted as $\mathcal{M}^*_{k^*}$. Finally, we find the best model within $\mathcal{M}^*_{k^*}$ which is the one with the optimal parameter values. In this case, there is no need to assume the existence of a "true" model or "true" parameters for the data in the preselected families. If the "true" model is in the preselected set of model families, then the best model family would most likely contain this "true" model, and so would the best model class, and our best model would most likely be the "true" model. This can be seen from the asymptotic results of the stochastic complexity introduced in the previous sections. On the other hand, the selection of good model families is precisely the place where human intuition and intelligence are indispensable.

In this thesis we study the application of the MDL principle in several statistical

modeling problems. This includes the principal components selection (Chapter 2), generalized linear model selection (Chapter 3), decision settlement of stochastic complexity estimation (Chapter 4), nonparametric testing hypothesis of homogeneity and general nonparametric histogram density estimation (Chapter 5 and 6). A number of new results have been obtained and will be presented in these chapters.

Chapter 2

Principal Components Selection by the Criterion of the Minimum Mean Difference of Complexity

2.1 Introduction

Principal components analysis, the first systematic account of which was developed by Hotelling (1933) as a data analytic technique, provides us with a method to deal with a large number of correlated variables, by which the dimension of the problem (that is, the number of variables) can be reduced without sacrificing too much of the information in the data. (For details see Kshirsagar (1972), Muirhead (1982) and Anderson (1984).)

In classical principal components selection, the criterion used to measure the lost information due to the reduction of the dimension of the observed variables is directly based on the covariance matrix Σ and can be applied only under the assumption of multinormality. When the observed variables do not follow a multinormal distribution the selection process, based on the covariance matrix Σ, does not provide a satisfactory explanation of the lost information. The theory of stochastic complexity or description length, which was developed in the works of Kolmogorov (1965), Rissanen (1989),

Wallace et. al. (1987), Barron and Cover (1991) etc. (see Chapter 1 for a more extended list of references), opens up a possibility to overcome this difficulty.

Let M_k denote a class of probability models, $M_k = \{P(x|\boldsymbol{\theta})\}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$ denotes a k-component parameter vector ranging over a subset Ω_k of the k-dimensional Euclidean space $\mathcal{R}^k$ with non-empty interior. This last condition is for convenience to ensure that the parameters are "free". Though the natural parameters are sometimes not free, one can always assume that by certain transformation some of the components can be eliminated and only the free ones remain. For a simple random sample $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$, drawn from a $q \times 1$ random vector $\underline{X}$ with probability function P, $P \in M_k$, the shortest code length for the description or complexity of the data is defined as

$$MDL(M_k) = \min_{\boldsymbol{\theta}} \left\{ L(\boldsymbol{\theta}) + \log \frac{1}{\prod_{i=1}^{n} P(\underline{x}_i)} \right\} \tag{2.1.1}$$

The nonnegative numbers $L(\boldsymbol{\theta})$ are assumed to satisfy Kraft's inequality $\sum_{\boldsymbol{\theta}^j} 2^{-L(\boldsymbol{\theta}^j)} \leq 1$, where $\boldsymbol{\theta}^j$ is the truncated vectors of $\boldsymbol{\theta}$ to the precision $\tau_i = 2^{-q_i}$, $i = 1, \cdots, k$, q_i are the number of fractional binary digits taken in the truncation, so that $L(\boldsymbol{\theta})$ corresponds to a prefix code C which describes the parameter vector $\boldsymbol{\theta}$.

In principal components analysis it is important that $MDL(M_k)$ is invariant under linear transformations of $\underline{X}$. The choice of a particular coordinate system, or units of measurement, is also very important as the principal components are meaningful only if all the variables are measured in the same units. If they are not, it is recommended that the analysis be performed on the standardized observations; in this case, questions of interpretation arise and the problems of inference are exceedingly complex, see Anderson (1963). For the sake of conciseness, we assume that all of the coordinates of $\underline{X}$ are measured in the same units and use orthogonal transformation so that the minimum description length is invariant under it.

Let $H = (\underline{h}_1, \underline{h}_2, \cdots, \underline{h}_q)$ be a $q \times q$ orthogonal matrix. If the random vector $\underline{X}$ is to be replaced by some variables in $\underline{h}_1^T \underline{X}, \underline{h}_2^T \underline{X}, \cdots, \underline{h}_q^T \underline{X}$, $q \leq n$, it is natural to consider one with the least difference in description length from the original data.

In this chapter we introduce the notion of an index of predictive power

$$IPP(V_{r,q}) = \min_{H_1 \in V_{r,q}} \left\{ \frac{k-k'}{2n} \log n + E_{\theta} \left(\log \frac{P_{H_1}(H_1^T \underline{X} \mid \theta')}{P(\underline{X} \mid \theta)} \right) \right\} \tag{2.1.2}$$

where $V_{r,q}$, $r \leq q$ is the Stiefel manifold defined by

$$V_{r,q} = \{q \times r \text{ matrix } H_1 \text{ satisfying } H_1^T H_1 = I_r\}. \tag{2.1.3}$$

$P_{H_1}(H_1^T \underline{x} \mid \theta')$ is the marginal probability function of $H_1^T \underline{X}$, with the parameter vector θ' ranging over a k'-dimensional Euclidean space with non-empty interior. $k' \leq k$ because the probability function $P(\underline{X} \mid \theta)$ is parameterized by the parameter vector $\theta \in \Omega^k$ having non-empty interior.

The components of $H_1^T \underline{X}$, for which H_1 achieves the right hand side of (2.1.2), are called the principal components of $\underline{X}$.

$IPP(V_{r,q})$ represents the expected difference of the complexity between the original variables and the principal components. If the parameter vector θ is known, we can find the value of H_1 for a prescribed number r (which indicates the number of principal components we will use) by solving the minimization problem of the right hand side of (2.1.2). For unknown θ we will show, by using the theory of the stochastic complexity, that the estimate $\widehat{IPP}(V_{r,q})$ obtained by substituting the MLE of θ in $IPP(V_{r,q})$ results in an optimal estimate of the expected difference of the complexity $IPP(V_{r,q})$. We can therefore use $\widehat{IPP}(V_{r,q})$ for finding principal components and regard the $IPP(V_{r,q})$ or $\widehat{IPP}(V_{r,q})$ as a criterion for principal components selection.

It will be shown that this criterion is equivalent with the classical one in which the covariance matrix Σ is used when the distribution is normal. The justification of the suggested selection process will be followed by a discussion of the principal components analysis for a class of ε-contaminated normal distributions, in which we show that the principal components change in a continuous manner with respect to ε in a small neighborhood of the "true" distribution.

2.2 Selection Principle for model $\{P(\underline{X} \mid \theta)\}$

Let $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$ be a simple random sample drawn from a $q \times 1$ random vector with probability function $P(\underline{X} \mid \theta)$. After Rissanen (1983) and (1978), the description length for $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$, per observation, is defined as

$$\frac{1}{n}L(X,\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log P(\underline{x}_i \mid \theta) + \frac{k}{2n}\log\left(\frac{2\pi e n}{k}\right) + \frac{k}{n}\log(\|\theta\|_{I(\theta)}) + O\left(\frac{\log k}{n}\right) \tag{2.2.1}$$

where $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$ and $\sum_{i=1}^{n}\log P(\underline{x}_i \mid \theta)$ denotes the log likelihood of the data for θ. $\|\theta\|_{I(\theta)} = \sqrt{\theta^T M(\theta)\theta/n}$ denotes the natural norm induced by the quadratic form associated with the $k \times k$ matrix $M(\theta)$ of the second derivatives of $-\sum_{i=1}^{n}\log P(\underline{x}_i \mid \theta)$ and $I(\theta) = M(\theta)/n$.

Expression (2.2.1), within a constant, is the negative logarithm of the joint probability of the data and the parameters. It can be obtained by optimizing the precision needed to express the parameters, and then using a universal prior distribution for the resulting integers, where the probability of integer n is proportional to $2^{-\log^* n}$. The function $\log^*$ is defined as $\log^* y = \log y + \log\log y + \cdots$, where only the positive terms are included in the sum. Notice that in deriving (2.2.1), Rissanen treated k as a variable, rather than a fixed number, so that both the estimation of the optimal θ as well as k could be based on (2.2.1). But here we prescribe k, the dimension of the parameter θ, as fixed by assuming a parametric family M_k, hence the term $O((\log k)/n)$ could be replaced by $O(1/n)$.

$I(\theta)$ is of order 1, provided that $-\sum_{i=1}^{n}\log P(\underline{x}_i \mid \theta)$ grows proportionally with n, as is the case normally . Then (2.2.1) can be expressed approximately as

$$\frac{1}{n}L(X,\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log P(\underline{x}_i \mid \theta) + \frac{k}{2n}\log n + O\left(\frac{1}{n}\right). \tag{2.2.2}$$

If the random vector $\underline{X}$ is not discrete, then no finite-length uniquely decodable codes exist. Nevertheless, quantization of the sample space of $\underline{X}$ does lead to outcomes that are finitely describable. Let $[\underline{X}]$ denote the quantization region that contains $\underline{X}$, $p(\underline{x} \mid \theta)$ denote the probability density function of $\underline{X}$ with respect to a known

σ-finite dominating measure $\nu(d\underline{x})$ which, for sake of simplicity, is chosen to be the usual Lebesgue measure , then $p(\underline{x} \mid \boldsymbol{\theta}) = \lim P([\underline{x}] \mid \boldsymbol{\theta})/\nu([\underline{x}])$ for almost every $\underline{x}$ (where the limit is taken for a refining sequence of quantization regions that generates $\underline{x}$). Consequently, $\log P([\underline{x}] \mid \boldsymbol{\theta}) \approx \log p(\underline{x} \mid \boldsymbol{\theta}) + \log \nu([\underline{x}])$ if the quantization is sufficiently fine. Using this approximation, for proper precision d used to express the sample values $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$, the description length for the sample, expressed per observation, is approximately given by

$$\frac{1}{n}L(X, \boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log p(\underline{x}_i \mid \boldsymbol{\theta}) + \frac{k}{2n}\log n - q \log d + O\left(\frac{1}{n}\right). \tag{2.2.3}$$

The term $-q \log d$ is the code length in coding the sample precision which can be regarded as fixed when the sample is given. The corrected description length then can be defined as

$$\frac{1}{n}L(X, \boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log p(\underline{x}_i \mid \boldsymbol{\theta}) + \frac{k}{2n}\log n + O\left(\frac{1}{n}\right). \tag{2.2.4}$$

This description length is minimized when $\boldsymbol{\theta}$ is replaced by its maximum likelihood estimate $\hat{\boldsymbol{\theta}}$.

For arbitrary $q \times r$ matrix $H_1 \in V_{r,q}$, the Stiefel manifold defined by (2.1.3) with $r < q$, we can find a $q \times (q - r)$ matrix H_2, so that $H_1 = (H_1, H_2)$ is an orthogonal matrix, then the density function of $H_1^T \underline{X}$, denoted by $p_{H_1}(H_1^T \underline{x} \mid \boldsymbol{\eta}(\boldsymbol{\theta}))$, can generally be obtained by integrating the density of $H^T \underline{X}$ with respect to $H_2^T \underline{x}$.

Often in stochastic models, the components in the parameter vector are not independent in the sense that they satisfy, either implicitly or explicitly, certain relationships among them. In the cases discussed, however, we assume that the dependent parameters have been eliminated after some transformation on them, and that the remaining k parameters range over the k-dimensional Euclidean space with non-empty interior.

By introducing some transformation $H_1 \in V_{r,q}$ of the random vector $\underline{X}$, we may impose more restrictions on the freedom of the parameter $\boldsymbol{\theta}$ because of the reduction of the dimension for $H_1^T \underline{X}$. Thus we may assume that the parameter vector is determined

by $\eta(\theta)$ in the marginal density $p_{H_1}(H_1^T \underline{x} \mid \eta(\theta))$ that ranges over a k'-dimensional Euclidean space with non-empty interior where $k' \leq k$, which is a function of θ and H_1 and in which the dependent structure has been removed by a certain transformation. For convenience we also assume that k' does not depend on the values in H_1 but is determined by r, the rank of H_1, as in the multinormal case. Under this assumption it can be shown that k' is an increasing function of r with $k'(q) = k$.

As in (2.2.4), the description length per observation for the data sequence $H_1^T \underline{x}_1$, $H_1^T \underline{x}_2, \cdots H_1^T \underline{x}_n$ is

$$\frac{1}{n} L(H_1^T X, \eta(\theta)) = -\frac{1}{n} \sum_{i=1}^{n} \log p_{H_1}(H_1^T \underline{x}_i \mid \eta(\theta)) + \frac{k'}{2n} \log n + O\left(\frac{1}{n}\right) \qquad (2.2.5)$$

To replace $\underline{X}$ by $H_1^T \underline{X}$ without too much loss of the information, we should select H_1 that minimizes the reduced description length

$$\Delta L = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{H_1}(H_1^T \underline{x}_i \mid \eta(\theta))}{p(\underline{x}_i \mid \theta)} + \frac{k - k'}{2n} \log n. \qquad (2.2.6)$$

This quantity is a random variable depending on $\underline{X}_i$, $i = 1, 2, \cdots, n$, for given θ and H_1. In order to understand the behavior of this minimization process, we replace (2.2.6) by its expected value and investigate the corresponding minimization problem.

Definition: The index of predictive power of $V_{r,q}$ (with respect to a model $p(\underline{x} \mid \theta)$ and sample size n) is defined by

$$IPP(V_{r,q}) = \min_{H_1 \in V_{r,q}} \left\{ \frac{k - k'}{2n} \log n + E_\theta \left(\log \frac{p_{H_1}(H_1^T \underline{X} \mid \eta(\theta))}{p(\underline{X} \mid \theta)} \right) \right\}. \qquad (2.2.7)$$

The components of $H_1^T \underline{X}$ are the corresponding principal components, where H_1 is the matrix minimizing (2.2.7).

$IPP(V_{r,q})$ is to be computed or estimated for every $r = 1, 2, \cdots, q-1$. If for some small m, $IPP(V_{m,q})$ is found small relative to $E_\theta((1/n)L(X, \theta))$ then most of the information in X is explained by $(H_1^{(m)})^T X$ (where the $m \times q$ matrix $H_1^{(m)}$ corresponds to $IPP(V_{m,q})$). $(H_1^{(m)})^T \underline{X}$ is proposed to be used as the first m principal components.

The solution H_1 is a function of $\boldsymbol{\theta}$. Since $(1/n)L(X, \hat{\boldsymbol{\theta}})$ is the description length of the least redundant encoding program for the observed $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$ it provides a natural estimate $\hat{H}_1$ of H_1 by replacing $\boldsymbol{\theta}$ with its maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Accordingly, the estimate of $IPP(V_{r,q})$, denoted as $\widehat{IPP}(V_{r,q})$, can be computed, and is identical with the minimization of the right hand side of (2.2.7) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. The minimum is achieved at $\hat{H}_1$.

In the light of the principle of the minimum description length (MDL) (which is a generalization of the maximum likelihood principle), the root of one optimal property of the estimate $\hat{\boldsymbol{\theta}}$, obtained by minimizing the description length (that happens to be the maximum likelihood estimate here), lies in the fact that the Kullback-Leibler distance between $p(X \mid \boldsymbol{\theta})$ and $p(X \mid \hat{\boldsymbol{\theta}})$ reaches asymptotically the minimum under certain mild smoothness conditions (see Theorem 1.7.1). This suggests that optimality might be achieved by using the corresponding estimates $\hat{H}_1$ and $\widehat{IPP}$. Unfortunately the large-sample distributional properties of this are still unclear and the construction of a test procedure for the validation of the selected principal components is intractably difficult. However, as we will see in Section 2.4, a universal test procedure can still be found with some desirable asymptotic properties based on the theory of the stochastic complexity.

If H_1 is a solution corresponding to $IPP(V_{r,q})$, then it is easy to see that for any $r \times r$ orthogonal matrix Q, H_1Q is also a solution corresponding to $IPP(V_{r,q})$. For the case of the normal distribution the principal components are uniquely defined except for a multiplicative $r \times r$ orthogonal matrix.

2.3 Principal Components in Normality Case

To illustrate the relationship between principal components selection by IPP and by the classic method we consider the case of the normal distribution. Let $\underline{X}$ be a $q \times 1$ random vector with multinormal distribution $N(\underline{\mu}, \Sigma)$ and $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$ be a simple random sample drawn from $\underline{X}$. For $H_1 \in V_{r,q}$, the distribution of $H_1^T \underline{X}$

is $N(H_1^T\underline{\mu}, H_1^T\Sigma H_1)$. The description length for X is

$$\begin{aligned}\frac{1}{n}L(X,\underline{\mu},\Sigma) &= -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\underline{x}_i-\underline{\mu})^T\Sigma^{-1}(\underline{x}_i-\underline{\mu})\right\}\right)\\ &\quad+\frac{q(q+3)}{4n}\log n+O\left(\frac{1}{n}\right)\\ &= \frac{q}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{2n}\sum_{i=1}^{n}tr\left(\Sigma^{-1}(\underline{x}_i-\underline{\mu})(\underline{x}_i-\underline{\mu})^T\right)\\ &\quad+\frac{q(q+3)}{4n}\log n+O\left(\frac{1}{n}\right). \qquad (2.3.1)\end{aligned}$$

The description length for H_1^TX is given by

$$\begin{aligned}&\tfrac{1}{n}L(H_1^TX,H_1^T\underline{\mu},H_1^T\Sigma H_1)\\ =\ &-\tfrac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{(2\pi)^{r/2}|H_1^T\Sigma H_1|^{1/2}}\exp\left\{-\frac{1}{2}(H_1^T\underline{x}_i-H_1^T\underline{\mu})^T\right.\right.\\ &\left.\left.(H_1^T\Sigma H_1)^{-1}(H_1^T\underline{x}_i-H_1^T\underline{\mu})\right\}\right)+\frac{r(r+3)}{4n}\log n+O\left(\frac{1}{n}\right)\\ =\ &\tfrac{r}{2}\log(2\pi)+\frac{1}{2}\log|H_1^T\Sigma H_1|+\frac{1}{2n}\sum_{i=1}^{n}tr\left((H_1^T\Sigma H_1)^{-1}H_1^T(\underline{x}_i-\underline{\mu})(\underline{x}_i-\underline{\mu})^TH_1\right)\\ &+\frac{r(r+3)}{4n}\log n+O\left(\frac{1}{n}\right). \qquad (2.3.2)\end{aligned}$$

After some simplification, the expected value of (2.2.6) is

$$\begin{aligned}&E\left(\frac{1}{n}L(X,\underline{\mu},\Sigma)-\frac{1}{n}L(H_1^TX,H_1^T\underline{\mu},H_1^T\Sigma H_1)\right)=\\ &\frac{q-r}{2}\log(2\pi)+\frac{1}{2}\log\frac{|\Sigma|}{|H_1^T\Sigma H_1|}+\frac{q}{2}-\frac{r}{2}+\frac{(q-r)(q+r+3)}{4n}\log n. \qquad (2.3.3)\end{aligned}$$

For the minimization of (2.3.3), the following Poincaré separation theorem from matrix theory is useful (see Chapter 1 in Rao (1973)).

Lemma 2.3.1 *(Poincaré): Let Σ be a $q\times q$ positive definite matrix whose eigenvalues are $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_q$, $H_1\in V_{r,q}$. Denote by $\lambda_1'\geq\lambda_2'\geq\cdots\geq\lambda_r'$ the eigenvalues of $H_1^T\Sigma H_1$, then*

$$\lambda_{q-r+i}\leq\lambda_i'\leq\lambda_i,\quad i=1,2,\cdots,r. \qquad (2.3.4)$$

By the lemma above, (2.3.3) is minimized at $|H_1^T \Sigma H_1| = \lambda_1 \lambda_2 \cdots \lambda_r$ and H_1 can be chosen as the unit and orthogonal eigenvectors of Σ corresponding to $\lambda_1, \lambda_2, \cdots, \lambda_r$ respectively. If Σ is replaced by its maximum likelihood estimate $S = (1/n)\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T$, where $\bar{\underline{x}} = (1/n)\sum_{i=1}^{n} \underline{x}_i$, then H_1 can be estimated by the first r orthogonal and unit eigenvectors of S corresponding to its first r eigenvalues. Let $l_1 \geq l_2 \geq \cdots \geq l_q$ be the eigenvalues of S, then the estimated $IPP(V_{r,q})$ is

$$\widehat{IPP}(V_{r,q}) = \frac{q-r}{2}\log(2\pi e) + \frac{(q-r)(q+r+3)}{4n}\log n + \frac{1}{2}\sum_{i=r+1}^{q}\log l_i, \quad r = 1, \cdots, q-1. \tag{2.3.5}$$

From the discussion above it follows that the principal components under criterion (2.2.7) are the same as the usual principal components for multinormal distribution.

2.4 Validation of the Principal Components

In Section 2.2 we deduced the index of predictive power as a descriptive measure for studying the dependence or correlational structure of multivariate samples drawn from a parametric model. Now the question arises whether the estimate of $IPP(V_{r,q})$ adequately describes the mean difference of complexity, how much confidence one can have in such principal-component estimation and how to construct hypothesis testing for principal-component selection with the associated confidence.

In classical principal components analysis, a number of large-sample distributional properties of the component coefficients and eigenvalues are derived. In addition to providing knowledge of the stability of these quantities through their variance-covariance structure, these asymptotic distributions allow the construction of tests of hypothesis and confidence intervals for the population component structure. The results have been summarized by Anderson (1984) and Muirhead (1982). Waternaux (1976) and Davis (1977) have studied the robustness of the principal-component distributions to nonnormality in the original observations. Waternaux concluded that tests or confidence intervals based on asymptotic distributional results could be seriously affected by nonnormality. Davis investigated the effects of nonnormality on

the hypothesis tests and confidence intervals for the eigenvalues and the eigenvectors, and gave conditions for the inferences to be conservative.

In our study, if the probability function of the assumed model are reasonably smooth, the principal-component coefficients matrix H_1 and hence $IPP(V_{r,q})$ will hold some smooth analytical properties such as continuity and differentiability. From the large-sample properties of the maximum likelihood estimate, it follows that the distribution of $\sqrt{n}\hat{\boldsymbol{\theta}}$ converges to the normal distribution with some mean $\boldsymbol{\theta}^*$ and covariance Σ^* under quite weak conditions, where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$ is the MLE computed from a sample $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n$. This Σ^* gives us an idea of the spread of $\hat{\boldsymbol{\theta}}$ at least for large n and we can construct a confidence interval for $\hat{\boldsymbol{\theta}}$ based on Σ^*. Accordingly, the confidence interval for H_1 and $IPP(V_{r,q})$ can be obtained.

In complexity theory, there is no need to assume the existence of any "true" parameter in the suggested model (see Rissanen (1989)). Frequently we fit parametric models of a certain kind to the observed data even though none of the models may capture all the major relevant features. We then solve the parameter estimate with which the induced model has the minimum description length or complexity among the suggested parametric models, and regard the induced model as the best model until a larger parametric model class is considered and/or another model is found with smaller complexity.

In practice, the bootstrap techniques are frequently used to provide Monte Carlo type estimates of covariance of $\hat{\boldsymbol{\theta}}$ and $\hat{H}_1$. We consider both the parametric and the nonparametric methods of bootstrap.

Nonparametric bootstrap: first we form an empirical distribution from the observations $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$ with $1/n$ the probability assigned to each $\underline{x}_i$. Then generate a sample X_1 of length n by sampling this empirical distribution n times and calculate a new estimate $\hat{\boldsymbol{\theta}}(X_1)$. Repeat the process N times and compute the means $\hat{\boldsymbol{\theta}}^* = (1/N)\sum_{i=1}^{N} \hat{\boldsymbol{\theta}}(X_i), \hat{H}_1^* = (1/N)\sum_{i=1}^{N} \hat{H}_1(X_i)$ and the desired covariance estimates

$$\hat{\Sigma}^* = (1/N)\sum_{i=1}^{N}(\hat{\boldsymbol{\theta}}(X_i) - \hat{\boldsymbol{\theta}}^*)(\hat{\boldsymbol{\theta}}(X_i) - \hat{\boldsymbol{\theta}}^*)^T,$$

$$\hat{\Sigma}^*_{H_1} = (1/N)\sum_{i=1}^{N} vec(\hat{H}_1(X_i) - \hat{H}_1^*)(vec(\hat{H}_1(X_i) - \hat{H}_1^*))^T.$$

In parametric bootstrap, we generate a new series of samples by sampling from the distribution $P(\underline{x} \mid \hat{\boldsymbol{\theta}})$ instead of the empirical distribution, where $\hat{\boldsymbol{\theta}}$ is the MLE from the original sample $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$.

When the index of predictive power (IPP) (2.2.7) is employed as a criterion for principal components selection, one need to deal with th problem that the IPP is a function of the unknown parameter $\boldsymbol{\theta}$. In order to turn it into an applicable data-based criterion we need to replace it by a suitable estimate $\widehat{IPP}$. Because of the fixed k in our assumed model class M_k, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is also the minimum description length estimate (refer to expression (2.2.4) and ignore the term $O(1/n)$). Consequently, it also provides us with an optimal density estimate $p(\underline{x} \mid \hat{\boldsymbol{\theta}})$ under the framework of the principle of minimum description length. Therefore $\widehat{IPP}$ may be obtained by performing the minimization (2.2.7) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. It is sometimes quite difficult to calculate the second term in the right hand side of (2.2.7) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, in which case we can approximate it by calculating the moment estimate $(1/n)\sum_{i=1}^{n} \log(p_{H_1}(\hat{H}_1^T \underline{x}_i \mid \eta(\hat{\boldsymbol{\theta}}))/p(\underline{x}_i \mid \hat{\boldsymbol{\theta}}))$.

As we have seen in Section 2.2, the selection of the principal components can be based on testing a series of hypotheses. Suppose the ratio of the lost information to the total information of the random vector $\underline{X}$, when using principal components instead of the original variables, is restricted to a prescribed value c, where $0 < c \leq 1$. If $(H_1^{(r)})^T \underline{X}$ are the first r principal components, where $H_1^{(r)}$ is obtained by minimizing (2.2.7), the lost information when using $(H_1^{(r)})^T \underline{X}$ is

$$J(H_1^{(r)}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left(\log \frac{p_{H_1^{(r)}}((H_1^{(r)})^T \underline{X} \mid \eta(\boldsymbol{\theta}))}{p(\underline{X} \mid \boldsymbol{\theta})} \right)$$

while the total information of $\underline{X}$ is the entropy

$$K(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}} \left(\log p(\underline{X} \mid \boldsymbol{\theta}) \right),$$

then our requirement becomes $J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) \leq c$. Notice that, the entropy with this definition may be negative and depends on the chosen coordinate system. For

simplicity, we assume that $K(\boldsymbol{\theta})$ is positive which can be achieved by choosing the coordinate system suitably, as in the case of multinormality. Using these notations, we can find how many principal components need to be used to meet such a requirement by testing the following hypotheses. First we test $A_1 : J(H_1^{(1)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) \leq c$ versus $B_1 : J(H_1^{(1)}, \boldsymbol{\theta})/K(\ \boldsymbol{\theta}) > c$. If A_1 is rejected we go on to test $A_2 : J(H_1^{(2)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) \leq c$ versus the corresponding B_2, etc. This procedure is continued until at some stage we can no longer reject the hypothesis $A_r : J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) \leq c$, where $r = 1, 2, \cdots, q$. The first r principal components $(\hat{H}_1^{(r)})^T \underline{X}$ can then be used to replace the original variables $\underline{X}$ in a statistical analysis without sacrificing more information than permitted. $r = q$ means that the dimension of the problem can not be reduced without violating the requirement that the proportion of the lost information is less than or equal to c.

Applying the idea of complexity, we propose a universal test statistic

$$\begin{aligned} T(X) &= \frac{\widehat{IPP}(V_{r,q})}{(1/n)E_{\boldsymbol{\theta}}(L(X,\boldsymbol{\theta}))\,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}} - c \\ &= \frac{\frac{k-k'}{2n}\log n + E_{\boldsymbol{\theta}}\left(\log \frac{p_{\hat{H}_1^{(r)}}((\hat{H}_1^{(r)})^T \underline{X}|\boldsymbol{\eta}(\boldsymbol{\theta}))}{p(\underline{X}|\boldsymbol{\theta})}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}{\frac{k}{2n}\log n - E_{\boldsymbol{\theta}}\left(\log p(\underline{X} \mid \boldsymbol{\theta})\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}} - c \end{aligned} \qquad (2.4.1)$$

for each hypothesis $A_r : J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) \leq c$ versus $B_r : J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) > c$, $r = 1, 2, \cdots, q$. To find a critical region for this test one may proceed in the usual manner to find a large-sample asymptotic distribution of $T(X)$, and then construct the critical region $T(X) > t$ where t is determined by the asymptotic distribution of $T(X)$ at $J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta}) = c$. However, we shall follow a different path here.

Consider the numerator of the first term of $T(X)$. It consists of two parts: the second part is the estimate of the lost information in the data when using the first r principal components, while the first part is a quantity measuring the reduction of the complexity of the model (in terms of the dimension of the parameter). Combining these two parts we get the estimate of the mean difference of the description lengths of X and $H_1^{(r)}X$, or the index of predictive power $IPP(V_{r,q})$ as derived in Section 2.2.

Similarly, the denominator of the first term of $T(X)$ consists of the estimate of the information per observation of the data X (the entropy) combined with the measure of the complexity of the proposed model $p(\underline{x} \mid \hat{\boldsymbol{\theta}})$ (in terms of the dimension of the parameter), and forms the mean minimum description length per observation of X. In this sense, the first term of $T(X)$ is not a simple estimate of $J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta})$; it estimates $J(H_1^{(r)}, \boldsymbol{\theta})/K(\boldsymbol{\theta})$ by attaching to it the penalty terms $(k/2n)\log n$ and $((k-k')/2n)\log n$, which are justified by providing the complexity and the reduction of the complexity respectively, for the model employed to give such an estimate.

Now we are in a position to suggest a critical region according to which the null hypothesis A_r is accepted if $T(X) \leq 0$, but rejected otherwise. With this test procedure the explicit knowledge of the distribution of the test statistics $T(X)$ is not required. Nor do we need to select the size of the test, or the type I error, for it is defined automatically from $P(T(X) > 0 \mid A_r)$, which depends on the two penalty terms and corresponds to an intuitively chosen significance level for some common sample sizes. Similar conclusion could be drawn for the type II error.

To clarify this, we abbreviate $J(H_1^{(r)}, \boldsymbol{\theta})$ as J, $K(\boldsymbol{\theta})$ as K and

$$T(X) = \frac{\frac{k-k'}{2n}\log n + \hat{J}}{\frac{k}{2n}\log n + \hat{K}} - c.$$

Also suppose that the MLE $\hat{\boldsymbol{\theta}}$ satisfies the central limit theorem at each interior point of Ω_k such that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(\underline{0}, \Sigma^*)$ in distribution and $(J/K)'$, the derivative with respect to $\boldsymbol{\theta}$, exists and $\neq 0$, then we know that

$$\sqrt{n}(\hat{J}/\hat{K} - J/K) \rightarrow N(0, ((J/K)')^T \Sigma^* (J/K)')$$

in distribution for each interior point of Ω_k.

From Section 2.2 we also know that k and k' are fixed after stating A_r and B_r. The size of the test is then

$$\begin{aligned} P(T(X) > 0 \mid A_r) &= P\left(\frac{\frac{k-k'}{2n}\log n + \hat{J}}{\frac{k}{2n}\log n + \hat{K}} - c > 0 \mid A_r\right) \\ &= P\left(\frac{\hat{J}}{\hat{K}} - \frac{J}{K} > \frac{\log n}{2n}\frac{k}{\hat{K}}\left(\frac{J}{K} - \frac{k-k'}{k}\right)\right. \end{aligned}$$

$$+\left(\frac{\log n}{2n}\frac{k}{\hat{K}}+1\right)\left(c-\frac{J}{K}\right) \mid A_r\Bigg)$$

which is determined by the last two terms and tends to 0 when $n \to \infty$ for $J/K < c$ under the assumptions above. For $J/K = c$ the expression above gives an asymptotic size of 0.5 for the test when $n \to \infty$. Hence the type I error could be fairly large for A_r versus B_r. But notice also that our test procedure is based on a series of hypotheses. For the fixed c our $J(H_1^{(r+1)}, \theta)/K(\theta)$ will be less than c (this can be seen from the definition of J,K and IPP), therefore the size of the test for A_{r+1} versus B_{r+1} will tend to 0 as $n \to \infty$. In short, for fixed c, the type I error of our principal components selection process (ignoring the mistake that the number of the principal components selected is one more or one less than the true number) is asymptotically 0 as $n \to \infty$.

By similar argument the type II error of $T(X)$ for A_r versus B_r,

$$\begin{aligned} P(T(X) \le 0 \mid B_r) &= P\left(\frac{\frac{k-k'}{2n}\log n + \hat{J}}{\frac{k}{2n}\log n + \hat{K}} - c \le 0 \mid B_r\right) \\ &= P\left(\frac{\hat{J}}{\hat{K}} - \frac{J}{K} \le \frac{\log n}{2n}\frac{k}{\hat{K}}\left(\frac{J}{K} - \frac{k-k'}{k}\right)\right. \\ &\quad \left. +\left(\frac{\log n}{2n}\frac{k}{\hat{K}}+1\right)\left(c-\frac{J}{K}\right) \mid B_r\right) \\ &\to 0 \quad \text{as } n \to \infty \end{aligned}$$

based on the same assumptions for deriving the size in the limit. This indicates some sort of asymptotic optimality of the power of the test statistics $T(X)$. In practice, both the size and the power could be found approximately by a Monte Carlo technique. In the case of multinormal distribution $T(X)$ is given by

$$T(X) = \frac{\frac{q-r}{2}\log(2\pi e) + \frac{(q-r)(q+r+3)}{4n}\log n + \frac{1}{2}\sum_{i=r+1}^{q}\log l_i}{\frac{q}{2}\log(2\pi e) + \frac{q(q+3)}{4n}\log n + \frac{1}{2}\sum_{i=1}^{q}\log l_i} - c. \qquad (2.4.2)$$

2.5 "ε-contaminated" Normal Random Vectors

In this section we discuss an important application of the principal components selection described above for the case of an "ε-contaminated" normal distribution, where the contaminating distribution is also normal.

Let $\underline{X}$ be distributed as an "ε-contaminated" normal distribution with density function

$$\begin{aligned} p(\underline{x},\Sigma,\sigma^2) &= (1-\varepsilon)\frac{1}{(2\pi)^{q/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}\underline{x}^T\Sigma^{-1}\underline{x}\right\} \\ &\quad +\varepsilon\frac{1}{(2\pi)^{q/2}\sigma^q}\exp\left\{-\frac{1}{2\sigma^2}\underline{x}^T\underline{x}\right\} \end{aligned} \tag{2.5.1}$$

where σ^2 is less than the minimum eigenvalue λ_q of Σ. Without loss of generality we assume that $\underline{X}$ has zero mean.

The marginal density function of $H_1^T\underline{X}$ is

$$\begin{aligned} p_{H_1}(H_1^T\underline{x}, H_1^T\Sigma H_1, \sigma^2) &= (1-\varepsilon)\frac{1}{(2\pi)^{r/2}|H_1^T\Sigma H_1|^{1/2}} \\ &\quad \exp\left\{-\frac{1}{2}(H_1^T\underline{x})^T(H_1^T\Sigma H_1)^{-1}(H_1^T\underline{x})\right\} \\ &\quad +\varepsilon\frac{1}{(2\pi)^{r/2}\sigma^r}\exp\left\{-\frac{1}{2\sigma^2}(H_1^T\underline{x})^T(H_1^T\underline{x})\right\}. \end{aligned} \tag{2.5.2}$$

In order to find the expected value of the difference in description lengths (2.2.6), we first compute

$$\begin{aligned} &E(\log p(\underline{X},\Sigma,\sigma^2)) \\ &= \log(1-\varepsilon) - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}E(tr(\Sigma^{-1}\underline{X}\underline{X}^T)) \\ &\quad +E\left(\log\left(1+\tfrac{\varepsilon}{1-\varepsilon}\tfrac{|\Sigma|^{1/2}}{\sigma^q}\exp\left\{-\tfrac{1}{2}tr\left(\left(\tfrac{1}{\sigma^2}I-\Sigma^{-1}\right)\underline{X}\underline{X}^T\right)\right\}\right)\right) \\ &= \log(1-\varepsilon) - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}tr(\Sigma^{-1}((1-\varepsilon)\Sigma+\varepsilon\sigma^2 I)) \\ &\quad +\textstyle\sum_{j=1}^{\infty}\frac{(-1)^{j-1}}{j}\left(\frac{\varepsilon}{1-\varepsilon}\right)^j\frac{|\Sigma|^{j/2}}{\sigma^{qj}}E\left(\exp\left\{-\tfrac{1}{2}tr\left(\left(\tfrac{1}{\sigma^2}I-\Sigma^{-1}\right)\underline{X}\underline{X}^T\right)\right\}\right) \\ &= \log(1-\varepsilon) - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{q}{2}(1-\varepsilon) - \frac{\varepsilon\sigma^2}{2}tr\Sigma^{-1} \\ &\quad +\varepsilon + \tfrac{1}{2}\tfrac{\varepsilon^2}{1-\varepsilon}\frac{|\Sigma|}{\sigma^q|2\Sigma-\sigma^2 I|^{1/2}} + O(\varepsilon^3). \end{aligned} \tag{2.5.3}$$

Similarly

$$\begin{aligned} &E(\log p_{H_1}(H_1^T\underline{X}, H_1^T\Sigma H_1, \sigma^2)) = \\ &\quad \log(1-\varepsilon) - \tfrac{r}{2}\log(2\pi) - \tfrac{1}{2}\log|H_1^T\Sigma H_1| - \tfrac{r}{2}(1-\varepsilon) \\ &\quad -\tfrac{\varepsilon\sigma^2}{2}tr(H_1^T\Sigma H_1)^{-1} + \varepsilon + \tfrac{1}{2}\tfrac{\varepsilon^2}{1-\varepsilon}\frac{|H_1^T\Sigma H_1|}{\sigma^r|2H_1^T\Sigma H_1-\sigma^2 I|^{1/2}} + O(\varepsilon^3). \end{aligned} \tag{2.5.4}$$

The expected difference of the description lengths is then

$$\begin{aligned} & E\left(\frac{1}{n}L(\underline{X},\Sigma,\sigma^2)-\frac{1}{n}L(H_1^T\underline{X},H_1^T\Sigma H_1,\sigma^2)\right) \\ &= \tfrac{(q-r)(q+r+1)}{4n}\log n+\tfrac{q-r}{2}\log(2\pi)+\tfrac{q-r}{2}(1-\varepsilon) \\ &\quad +\tfrac{1}{2}\log\tfrac{|\Sigma|}{|H_1^T\Sigma H_1|}+\tfrac{\varepsilon\sigma^2}{2}tr(\Sigma^{-1}-(H_1^T\Sigma H_1)^{-1}) \\ &\quad +\tfrac{1}{2}\tfrac{\varepsilon^2}{1-\varepsilon}\left(\tfrac{|H_1^T\Sigma H_1|}{\sigma^r|2H_1^T\Sigma H_1-\sigma^2 I|^{1/2}}-\tfrac{|\Sigma|}{\sigma^q|2\Sigma-\sigma^2 I|^{1/2}}\right)+O(\varepsilon^3). \end{aligned} \tag{2.5.5}$$

The minimization of (2.5.5) is equivalent to the minimization of

$$\begin{aligned} F(H_1) &= \log\frac{|\Sigma|}{|H_1^T\Sigma H_1|}+\varepsilon\sigma^2 tr(\Sigma^{-1}-(H_1^T\Sigma H_1)^{-1}) \\ &\quad +\frac{\varepsilon^2}{1-\varepsilon}\frac{|H_1^T\Sigma H_1|}{\sigma^r|2H_1^T\Sigma H_1-\sigma^2 I|^{1/2}} \end{aligned} \tag{2.5.6}$$

when ε is sufficiently small. Using the notation of Lemma 2.3.1, we obtain

$$\begin{aligned} F(H_1) &= \sum_{i=1}^{q}\left(\log\lambda_i+\frac{\varepsilon\sigma^2}{\lambda_i}\right)-\sum_{i=1}^{r}\left(\log\lambda_i'+\frac{\varepsilon\sigma^2}{\lambda_i'}\right) \\ &\quad +\frac{\varepsilon^2}{1-\varepsilon}\frac{\prod_{i=1}^{r}\lambda_i'}{\sigma^r\prod_{i=1}^{r}(2\lambda_i'-\sigma^2)^{1/2}} \end{aligned} \tag{2.5.7}$$

and

$$\begin{aligned} \frac{\partial F(H_1)}{\partial\lambda_i'} &= -\frac{1}{\lambda_i'}+\frac{\varepsilon\sigma^2}{\lambda_i'^2}+\frac{\varepsilon^2}{1-\varepsilon}\frac{\prod_{j\neq i}\lambda_j'(\lambda_i'-\sigma^2)}{\sigma^r\prod_{j=1}^{r}(2\lambda_j'-\sigma^2)^{1/2}(2\lambda_i'-\sigma^2)} \\ &\leq -\frac{1}{\lambda_i'}+\frac{\varepsilon}{\lambda_i'}+\frac{\varepsilon^2}{1-\varepsilon}\frac{\prod_{j\neq i}\lambda_j'(\lambda_i'-\sigma^2)}{\sigma^r\prod_{j=1}^{r}(2\lambda_j'-\sigma^2)^{1/2}(2\lambda_i'-\sigma^2)} \\ &= \frac{-(1-\varepsilon)^2\sigma^r\prod_{j=1}^{r}(2\lambda_j'-\sigma^2)^{1/2}(2\lambda_i'-\sigma^2)+\varepsilon^2\prod_{j=1}^{r}\lambda_j'(\lambda_i'-\sigma^2)}{(1-\varepsilon)\lambda_i'\sigma^r\prod_{j=1}^{r}(2\lambda_j'-\sigma^2)^{1/2}(2\lambda_i'-\sigma^2)} \end{aligned} \tag{2.5.8}$$

by the condition $\sigma^2<\lambda_q$, $\varepsilon<1$ and $\lambda_{q-r+i}\leq\lambda_i'\leq\lambda_i$, $i=1,2,\cdots,r$. (2.5.8) ≤ 0 if

$$(1-\varepsilon)^2\sigma(2\lambda_j'-\sigma^2)^{1/2}\geq\varepsilon^2\lambda_j', \quad j=1,2,\cdots r \tag{2.5.9}$$

and (2.5.9) satisfies if $\lambda_q\leq\tau\sigma^2$, where $\tau=1/(1-\sqrt{1-(\varepsilon/(1-\varepsilon))^4})$. Hence we conclude that if ε is sufficiently small, $F(H_1)$ is minimized when $\lambda_i'=\lambda_i$, $i=1,2,\cdots,r$ and so is (2.5.5).

The first r principal components coefficients, i.e. the matrix H_1, are therefore the r orthogonal eigenvectors of unit length of Σ corresponding to the first r eigenvalues of Σ respectively.

Actually, to satisfy condition (2.5.9), the requirement for ϵ need not be too stringent. Table 2.1 below shows some ϵ and τ values and the corresponding values of ϵ^3 which indicate the precision in (2.5.3) – (2.5.5).

Table 2.1:

ϵ	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
τ	260642	13122	2062	511.5	161.5	58.8	23.3	9.6	3.9
ϵ^3	0.0001	0.001	0.003	0.008	0.016	0.027	0.043	0.064	0.091

The estimation of H_1 can also be done through the maximum likelihood estimate of Σ. Within a constant, the log likelihood function for $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)$ is

$$-\frac{n}{2}\log|\Sigma| - \frac{n}{2}tr(\Sigma^{-1}S) + \sum_{i=1}^{n}\log\left(1 + \frac{\epsilon}{1-\epsilon}\frac{|\Sigma|^{1/2}}{\sigma^q}\exp\left\{-\frac{1}{2}\underline{x}_i^T\left(\frac{1}{\sigma^2}I - \Sigma^{-1}\right)\underline{x}_i\right\}\right) \tag{2.5.10}$$

where $S = (1/n)\sum_{i=1}^{n}\underline{x}_i\underline{x}_i^T$.

If ϵ is small enough, (2.5.10) can be approximately expressed as

$$\begin{aligned}\ell(\Sigma, \sigma^2; X) &= -\frac{n}{2}\log|\Sigma| - \frac{n}{2}tr(\Sigma^{-1}S) + \frac{\epsilon}{1-\epsilon}\frac{n|\Sigma|^{1/2}}{\sigma^q}\left(1 - \frac{1}{2}tr\left(\left(\frac{1}{\sigma^2}I - \Sigma^{-1}\right)S\right)\right) \\ &= -\frac{n}{2}\log|S| + \frac{n}{2}\sum_{i=1}^{q}\log\delta_i - \frac{n}{2}\sum_{i=1}^{q}\delta_i \\ &\quad + \frac{\epsilon}{1-\epsilon}\frac{n|S|^{1/2}}{\sigma^q}\left(\prod_{i=1}^{q}\delta_i^{-1/2}\right)\left(1 - \frac{1}{2\sigma^2}trS + \frac{1}{2}\sum_{i=1}^{q}\delta_i\right) \end{aligned} \tag{2.5.11}$$

where $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_q$ are the eigenvalues of $S^{1/2}\Sigma^{-1}S^{1/2}$.

Because σ^2 is less important than Σ we use $\hat{\sigma}^2$, which equals to the smallest eigenvalue of $(1/2)S$, to replace σ^2 in (2.5.11). This way we could guarantee that the

condition $\sigma^2 < \lambda_q$ holds in most cases. From $\frac{\partial \ell(\Sigma, \hat{\sigma}^2; X)}{\partial \delta_i} = 0$, $i = 1, 2, \cdots, q$, we obtain

$$\frac{1}{\delta_i} - 1 - \frac{\varepsilon}{1-\varepsilon} \frac{n|S|^{1/2}}{\hat{\sigma}^q} \left(\prod_{j=1}^{q} \delta_j^{-1/2} \right) \delta_j^{-1} \left(1 - \frac{1}{2\hat{\sigma}^2} trS + \frac{1}{2} \sum_{j=1}^{q} \delta_j \right)$$
$$+ \frac{\varepsilon}{1-\varepsilon} \frac{n|S|^{1/2}}{\hat{\sigma}^q} \left(\prod_{j=1}^{q} \delta_j^{-1/2} \right) = 0, \quad i = 1, 2, \cdots, q. \qquad (2.5.12)$$

Solving this, we get

$$\delta_i = \frac{1 - \frac{\varepsilon}{1-\varepsilon} \frac{n|S|^{1/2}}{\hat{\sigma}^q} \left(\prod_{j=1}^{q} \delta_j^{-1/2} \right) \left(1 - \frac{1}{2\hat{\sigma}^2} trS + \frac{1}{2} \sum_{j=1}^{q} \delta_j \right)}{1 - \frac{\varepsilon}{1-\varepsilon} \frac{n|S|^{1/2}}{\hat{\sigma}^q} \left(\prod_{j=1}^{q} \delta_j^{-1/2} \right)}, \quad i = 1, 2, \cdots, q. \quad (2.5.13)$$

The solutions δ_i's of (2.5.13) for $i = 1, 2, \cdots, q$ are identical and the common value is the solution of

$$\delta = \frac{1 - \frac{\varepsilon}{1-\varepsilon} \frac{n|S|^{1/2}}{\hat{\sigma}^q} \delta^{-q/2} \left(1 - \frac{1}{2\hat{\sigma}^2} trS + \frac{q}{2} \delta \right)}{1 - \frac{\varepsilon}{1-\varepsilon} \frac{n|S|^{1/2}}{\hat{\sigma}^q} \delta^{-q/2}} \qquad (2.5.14)$$

which is equivalent to a polynomial equation for δ. Noticing that

$$\frac{\partial^2 \ell(\Sigma, \hat{\sigma}^2; X)}{\partial \delta_i \partial \delta_j} = -\frac{1}{\delta_i^2} \xi_{ij} + O(\varepsilon)$$

for $i, j = 1, 2, \cdots, q$, where $\xi_{ij} = 1$ for $i = j$ and 0 otherwise, the MLE's of δ_i in (2.5.11) exist and are unique if ε is small enough. If we denote the solution by $\hat{\delta}_i = 1 + \nu$, then ν is a continuous function of ε and $\lim_{\varepsilon \to 0} \nu = 0$, and the MLE of Σ is $\hat{\Sigma} = (1 + \nu)^{-1} S$.

The H_1 which minimizes (2.5.5) can be estimated by the first r orthogonal eigenvectors of unit length of $\hat{\Sigma}$. When $\varepsilon \to 0$, the estimate of H_1 tends to that of the solution which minimizes (2.3.3). This result is summarized in the following theorem.

Theorem 2.5.1 *If $H_1^{(\varepsilon)}$ is the solution corresponding to $IPP(V_{r,q})$ for the "ε-contaminated" normal distribution (2.5.1) and H_1 is the solution corresponding to the normal distribution $N(\underline{0}, \Sigma)$, then their MLE estimates, denoted by $\hat{H}_1^{(\varepsilon)}$ and $\hat{H}_1$, are approximately the first r orthogonal eigenvectors of unit length of $(1 + \nu)^{-1} S$ and S respectively. Furthermore, $\lim_{\varepsilon \to 0} \hat{H}_1^{(\varepsilon)} = \hat{H}_1$. For the "$\varepsilon$-contaminated" normal distribution (2.5.1), $IPP(V_{r,q})$ is estimated by*

$$\widehat{IPP}(V_{r,q}) = \frac{(q-r)(q+r+1)}{4n} \log n + \frac{(q-r)}{2} \log(2\pi e^{1-\varepsilon})$$

$$+\frac{1}{2}\sum_{i=r+1}^{q}\log\hat{\lambda}_i+\frac{\varepsilon\hat{\sigma}^2}{2}\sum_{i=r+1}^{q}\frac{1}{\hat{\lambda}_i}$$
$$+\frac{1}{2}\frac{\varepsilon^2}{1-\varepsilon}\left(\frac{\prod_{i=1}^{r}\hat{\lambda}_i}{\hat{\sigma}^r\prod_{i=1}^{r}(2\hat{\lambda}_i-\hat{\sigma}^2)^{1/2}}-\frac{\prod_{i=1}^{q}\hat{\lambda}_i}{\hat{\sigma}^q\prod_{i=1}^{q}(2\hat{\lambda}_i-\hat{\sigma}^2)^{1/2}}\right) \quad (2.5.15)$$

where $\hat{\sigma}^2$ *equals to the smallest eigenvalue of* $(1/2)S$ *and* $\hat{\lambda}_1>\hat{\lambda}_2>\cdots>\hat{\lambda}_q$ *are the eigenvalues of* $\hat{\Sigma}=(1+\nu)^{-1}S$.

Remark 1: In the discussion above, it is necessary to assume that $\sigma^2<\lambda_q$. This assumption is a reasonable one if, as usually is the case, the contaminating observations take up only a small part of the sample and have a smaller variation than the uncontaminated part.

Remark 2: The above discussion can be extended to more general ε-contaminated normal models

$$(1-\varepsilon)\frac{1}{(2\pi)^{q/2}|\Sigma_1|^{1/2}}\exp\left\{-\frac{1}{2}tr(\Sigma_1^{-1}(\underline{X}-\underline{\mu}_1)(\underline{X}-\underline{\mu}_1)^T)\right\}$$
$$+\varepsilon\frac{1}{(2\pi)^{q/2}|\Sigma_2|^{1/2}}\exp\left\{-\frac{1}{2}tr(\Sigma_2^{-1}(\underline{X}-\underline{\mu}_2)(\underline{X}-\underline{\mu}_2)^T)\right\} \quad (2.5.16)$$

where $\underline{\mu}_1,\underline{\mu}_2,\Sigma_1,\Sigma_2$ and ε all are unknown.

After some calculation, it can be shown that finding the solution of H_1 for model (2.5.16) is asymptotically equivalent to finding $H_1\in V_{r,q}$ which minimizes

$$\begin{aligned}F(H_1) &= -\log|H_1^T\Sigma_1H_1|-\frac{\varepsilon}{2}tr((H_1^T\Sigma_1H_1)^{-1}(H_1^T\Sigma_2H_1))\\ &\quad-\frac{\varepsilon}{2}(H_1^T(\underline{\mu}_2-\underline{\mu}_1))^T(H_1^T\Sigma_1H_1)^{-1}(H_1^T(\underline{\mu}_2-\underline{\mu}_1))\\ &\quad+\frac{1}{2}\frac{\varepsilon^2}{1-\varepsilon}\frac{|H_1^T\Sigma_1H_1|}{|H_1^T\Sigma_2H_1|^{1/2}|H_1^T(2\Sigma_1-\Sigma_2)H_1|^{1/2}}\\ &\quad\exp\left\{\frac{1}{2}(H_1^T(\underline{\mu}_2-\underline{\mu}_1))^T(H_1^T\Sigma_1H_1)^{-1}(H_1^T(\underline{\mu}_2-\underline{\mu}_1))\right\}. \quad (2.5.17)\end{aligned}$$

With $\underline{\mu}_i,\Sigma_i$ and ε being fixed, this problem can be solved by an appropriate numerical method. The remaining question is to find the maximum likelihood estimates of $\underline{\mu}_i$, Σ_i and ε. Fortunately good results already exist and the computation for MLE's is fairly routine. A detailed account is given in Everitt and Hand (1981), and Wolfe

(1970), where it is shown that the likelihood equations for finite mixtures are the weighted averages of the likelihood equations arising from each component density in the mixture separately. The weights are the posterior probabilities of an observation arising from a particular component. Generally, the equations must be solved by some type of iterative procedure, the most useful being the E-M algorithm of Dempster et al. (1977).

Finally, we present the results of a simulation study which illustrate the theoretical results for the ε-contaminated normal model.

An artificial 6×1 random vector $\underline{X} = (X_1, X_2, \cdots, X_6)$ was generated with joint density (2.5.1), where $n = 115$, $\varepsilon = 15/115 \approx 0.13$, $\sigma^2 = 0.4$ and

$$\Sigma = \begin{pmatrix} 3.5536 & 1.4463 & 0.3891 & 0.8356 & -0.5784 & -0.2497 \\ 1.4463 & 3.5536 & -0.3892 & -0.8357 & 0.5783 & 0.2496 \\ 0.3891 & -0.3892 & 1.4717 & -1.8686 & 0.2738 & -1.0102 \\ 0.8356 & -0.8357 & -1.8686 & 8.0128 & -2.1330 & 2.2440 \\ -0.5784 & 0.5783 & 0.2738 & -2.1330 & 1.5730 & -0.4447 \\ -0.2497 & 0.2496 & -1.0102 & 2.2440 & -0.4447 & 1.8351 \end{pmatrix}.$$

The results are listed in Table 2.2. These are to be compared with the classical principal components and with two kinds of robust principal components obtained through minimum volume ellipsoid covariance matrix estimate (MVE) and through weighting on Mahalanobis distance (Table 2.3, 2.4 and 2.5, respectively). The results are quite close to each other (note that the last two rows in Table 2.3 and 2.5 are exactly the same; it is so because the weights are all 1 in this case). For details about the robust estimates of covariance matrix, see Lopuhaa and Rousseeuw (1991), Rousseeuw (1991), Rousseeuw and van Zomeren (1990).

Table 2.2: Principal Components based on model (2.5.1)

eigenvalues	10.0	5.0	3.0	0.8	0.7	0.5
	$\underline{h}_1$	$\underline{h}_2$	$\underline{h}_3$	$\underline{h}_4$	$\underline{h}_5$	$\underline{h}_6$
coefficients	-0.0851	-0.6854	-0.6106	0.0	-0.3238	-0.4074
of principal	0.0851	-0.6854	0.6106	0.0	0.3238	0.4074
components	0.2083	0.0	-0.4986	0.0	-0.2644	0.9980
	-0.8126	0.0	-0.0516	0.2981	0.0820	0.2878
	0.2377	0.0	0.3302	0.5987	-0.5254	-0.0842
	-0.2568	0.0	0.4687	-0.3891	-0.7457	0.0910
IPP	7.8366	5.6859	3.8018	2.4722	1.1784	0.0
*	73.86%	53.59%	35.83%	23.30%	11.11%	0.0%

Table 2.3: Principal Components based on the Sample Covariance Matrix

eigenvalues	9.3481	4.9491	2.7464	0.7626	0.7185	0.4135
	$\underline{h}_1$	$\underline{h}_2$	$\underline{h}_3$	$\underline{h}_4$	$\underline{h}_5$	$\underline{h}_6$
coefficients	0.2199	0.7591	-0.4258	-0.3136	0.1286	0.2813
of principal	0.0794	0.6072	0.6299	0.3511	-0.1236	-0.2994
components	-0.2421	0.1064	-0.4177	-0.0919	0.0650	-0.8619
	0.8643	-0.1756	-0.0907	-0.0718	-0.3876	-0.2421
	-0.2524	0.0299	0.3207	-0.7712	-0.4864	-0.0353
	0.2757	-0.1093	0.3693	-0.4124	0.7597	-0.1687
**	7.8479	5.6914	3.8025	2.5188	1.1814	0.0
***	73.96%	53.64%	35.84%	23.74%	11.13%	0.0%

Table 2.4: Principal Components based on the MVE Covariance Matrix Estimate

eigenvalues	8.5540	4.4575	2.2117	0.7856	0.6967	0.3811
	$\underline{h}_1$	$\underline{h}_2$	$\underline{h}_3$	$\underline{h}_4$	$\underline{h}_5$	$\underline{h}_6$
coefficients	0.2360	0.7579	-0.4147	-0.3299	-0.0787	-0.2880
of principal	0.0260	0.6102	0.6420	0.3876	0.0196	0.2533
components	-0.2033	0.1256	-0.4033	-0.0778	-0.1326	0.8699
	0.8704	-0.1383	-0.0354	0.0231	0.3869	0.2680
	-0.2779	0.0639	0.2811	-0.6436	0.6451	0.0969
	0.2598	-0.1196	0.4159	-0.5658	-0.6403	0.1226
**	7.8455	5.6924	3.8037	2.5285	1.1839	0.0
***	73.94%	53.65%	35.85%	23.83%	11.16%	0.0%

Table 2.5: Principal Components based on the Weighted Covariance Matrix Estimate

eigenvalues	9.2668	4.9060	2.7225	0.7559	0.7123	0.4099
	$\underline{h}_1$	$\underline{h}_2$	$\underline{h}_3$	$\underline{h}_4$	$\underline{h}_5$	$\underline{h}_6$
coefficients	0.2199	0.7591	-0.4258	-0.3136	0.1286	0.2813
of principal	0.0794	0.6072	0.6299	0.3511	-0.1236	-0.2994
components	-0.2421	0.1064	-0.4177	-0.0919	0.0650	-0.8619
	0.8643	-0.1756	-0.0907	-0.0718	-0.3876	-0.2421
	-0.2524	0.0299	0.3207	-0.7712	-0.4864	-0.0353
	0.2757	-0.1093	0.3693	-0.4124	0.7597	-0.1687
**	7.8479	5.6914	3.8025	2.5188	1.1814	0.0
***	73.96%	53.64%	35.84%	23.74%	11.13%	0.0%

*: The ratio of IPP to expected description length per observation which is 10.6104 computed using (2.5.3).

**: The expected difference of the description lengths (2.5.5).

***: The ratio of ** to 10.6104, the expected description length per observation.

Chapter 3

Generalized Linear Model Selection by Predictive Least Quasi-deviance Criterion

3.1 Introduction

Several criteria are available in the literature of model selection. See e.g. Akaike, 1973,1974; Efron, 1983,1986; Jaynes, 1957,1982,1985; Mallows, 1973; Schwarz, 1978; Shao, 1993; Shibata, 1981; and Stone, 1974. In addition we have seen in Chapter 1 the development of two new general approaches to problem of statistical inference: prequential analysis (Dawid, 1984, 1991a, 1991b) and stochastic complexity (Solomonoff, 1978; Rissanen, 1978, 1986a, 1987, 1989). The former approach is based on the idea that one of the purposes of statistics is to make sequential probability forecasts for future observations, and statistical methods should be assessed by means of the validity of the predictions that flow from them. Whereas in the latter approach a statistical model is characterized in terms of the length of a coded message needed to transmit the data, and the empirical assessment of the models are based on these code lengths. These two approaches are particularly well suited to model selection in the sense that both methods compare different models by their accumulated prediction

errors although with different interpretations. (For relationship between these two approaches see Dawid (1992).)

One of the basic notions in Rissanen's approach is the concept of predictive stochastic complexity, and in association with it a model selection procedure called the predictive minimum description length principle (Rissanen, 1986a, 1987, 1989). This principle, unlike for example the maximum likelihood method, permits optimal identification of the values as well as the number of the parameters. When restricted to Gaussian regression models, the predictive minimum description length principle gives rise to the predictive least squares principle (Rissanen, 1986b). Whereas the usual least squares technique minimizes the sum of squared fitting errors (residuals), the predictive least squares principle minimizes the accumulated squared prediction errors of the observations. Its minimization criterion contains the sum of the squared "honest" prediction errors (by "honest" we mean that only past data are used to identify the parameters in the model) which is shown to be an approximation of the predictive stochastic complexity of the data except for a multiplicative constant. (For a discussion of the predictive least squares principle see Hannan et al (1989), Hemerly and Davis (1989), Speed and Yu (1993), Wax (1988) and Wei (1992).)

In this chapter we propose a criterion for generalized linear model selection based on the predictive minimum description length principle and the idea of prequential analysis, as well as on some results in the theory of quasi-likelihood functions (McCullagh and Nelder, 1989; Wedderburn,1974).

Suppose the components of the response n-vector $Y = (y_1, \cdots, y_n)^T$ are independent variables with mean vector $\mu = (\mu_1, \cdots, \mu_n)^T$ and each with a covariance $\sigma^2 V_i(\mu_i)$, where the scalar σ^2 is a constant of probably unknown value and $V_i(\cdot)$ is a known positive function. It is assumed that the $p \times 1$ vector β is the parameter of interest and it is connected with μ through a generalized linear regression equation $g(\mu) = X\beta$, where $X = (x_1, \cdots, x_n)^T$ is an $n \times p$ matrix of the observed $p \times 1$ covariate vector x (the predictors) and $g(\cdot)$ is a link function.

Now we consider the problem of selecting a model (i.e., a regressor $X\beta$) that

minimizes the sum of the "honest" predictive quasi-deviance

$$\sum_{i=1}^{n} \int_{\hat{\mu}_i(i)}^{y_i} \frac{y_i - t}{V_i(t)} dt, \tag{3.1.1}$$

where $\hat{\mu}_i(i)$ is the estimate of the mean of y_i based on the first $i-1$ response values and the corresponding values of the employed predictors through the usual maximum quasi-likelihood method.

If the likelihood of Y takes the form

$$\exp\left\{\sigma^{-2}(Y^T\theta - b(\theta)) + c(Y,\sigma)\right\} \tag{3.1.2}$$

for suitably chosen functions $b(\theta)$ of the n-dimensional parameter θ and $c(Y,\sigma)$, it will be seen that (3.1.1) is the predictive stochastic complexity of Y relative to this model, or equivalently the negative prequential log-likelihood of the model on Y, all being in agreement up to an (data-dependent) irrelevant quantity for the model selection. Therefore, the model selection based on (3.1.1) can actually be interpreted as an extension of both the stochastic complexity approach and the prequential analysis.

All of our results are obtained for a class of finite dimensional models, in contrast with those discussed in Shibata (1983a, 1983b), Breiman and Freedman (1983) etc., where infinite dimensional models are also considered.

The main result of this chapter is to show that by minimizing (3.1.1) over a sufficiently large class of models, the probability of selecting the right model converges to 1, and the selected model converges to the optimal model in expectation. Here the optimal model is defined to be the correct model $g(\mu) = X\beta$ relative to a link function $g(\cdot)$, which has the smallest dimension among all the available ones. By using a resampling technique the proposed Monte Carlo predictive least quasi-deviance method is shown through a simulation study to have fairly strong power to enhance the efficiency in selecting the optimal model.

3.2 Model Selection and the "Honest" Predictive Error

Using the notations and assumptions described above, we consider the generalized linear regression model in which the systematic component is $g(\mu) = X\beta$.

Suppose the covariate vector x contains all the possible explanatory variables available, and suppose also that the dimension of x, denoted as p, is finite, which is usually the case in the practical situation. By including all these variables in the regression model we make use of all the information of the data. This, however, may be inefficient because some of the components of β may equal to zero, so that the corresponding explanatory variables are superfluous. The question arises then: how to choose the explanatory variables so that the resulting regression model is correct as well as efficient?

If some of the components of β are zero, a more compact model might be

$$g(\mu_\alpha) = X_\alpha \beta_\alpha, \tag{3.2.1}$$

where α is a subset of size p_α of $\{1, \cdots, p\}$, $p_\alpha \leq p$, β_α is a $p_\alpha \times 1$ vector containing the components of β indexed by the integers in α, $X_\alpha = (x_{1\alpha}, \cdots, x_{n\alpha})^T$ containing the columns of X indexed also by the integers in α, and $\mu_\alpha = (\mu_{\alpha 1}, \cdots, \mu_{\alpha n})$ is the assumed mean of Y under this model.

There are in total $2^p - 1$ possible different models of the form (3.2.1) each of which corresponds to a subset α and is denoted by $\mathcal{M}_\alpha$. The dimension (or size) of $\mathcal{M}_\alpha$ is defined to be p_α, the dimension of the vector β_α. Let $\mathcal{A}$ denote all nonempty subsets of $\{1, \cdots, p\}$. Following Shao (1993) the class of models $\mathcal{M}_\alpha$ can be grouped into two categories:

Category I: At least one non-zero component of β is not in β_α;

Category II: β_α contains all non-zero components of β.

Clearly, the models in Category I are incorrect models and the models in Category II are correct, but possibly inefficient, due to their large size. Among the $2^p - 1$

models an optimal model, denoted by $\mathcal{M}_*$, is defined to be the model in Category II with the smallest dimension. Note that a model is meant to be optimal relative to a fixed link function $g(\cdot)$. Generally the optimal model does not have to be unique (e.g. an essential explanatory variable is included twice in x), but if we assume that the components of x are linearly independent (i.e. if there exists a $p \times 1$ vector b such that $x^T b = 0$ then $b \equiv 0$), the optimal model is unique relative to the fixed $g(\cdot)$ and identical to the model in Category II with the smallest dimension.

Either by properly designing the experiment or by an appropriate transformation of the explanatory variable or both the linear independence of the components of x can usually be achieved. It is therefore meaningful to perform the model selection based on the above classification of the models in $\mathcal{M}_\alpha$, which is equivalent to the problem of variable (predictor) selection.

For other model selection procedures like Akaike's AIC or Schwarz's BIC, which is so formulated that each employed model is indexed by its dimension, refer to Akaike (1974), Nishii (1984) and Schwarz (1978) for detail.

Under the specified conditions and assumptions for Y the log quasi-likelihood function for Y is given by

$$Q(\mu; Y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V_i(t)} dt$$

and the quasi-deviance function for Y is

$$D(Y; \mu) = -2\sigma^2 Q(\mu; Y) = 2 \sum_{i=1}^{n} \int_{\mu_i}^{y_i} \frac{y_i - t}{V_i(t)} dt.$$

provided that the summands exist.

For generalized linear regression models, the quasi-likelihood method. suggested by Wedderburn (1974), behaves like the maximum likelihood method. The difference is that the assumptions of the former method concern only the first and the second moments and some additional regularity conditions relating to the regression equation. This relationship may be understood by looking at the least squares method and the maximum likelihood method in a linear regression model. For a detailed

description of the quasi-likelihood method, see Wedderburn (1974) and McCullagh and Nelder (Chapter 9, 1989).

Let us suppose that the components of Y are ordered and let $Y_i = (y_1, \cdots, y_{i-1})^T$ and $X^{(i)} = (x_1, \cdots, x_{i-1})^T$ be $(i-1) \times p$ matrices comprising rows of covariates (predictors) corresponding to the response variables, $i = 2, \cdots, n$. $X = (x_1, \cdots, x_n)^T$ is an $n \times p$ covariates matrix.

Under the proposed model $\mathcal{M}_\alpha$ of the form (3.2.1), the maximum quasi-likelihood estimate $\hat{\beta}_\alpha(i)$ of β_α, based on the first $i-1$ response values Y_i and the corresponding $X^{(i)}$, satisfies the estimating equations $U(\hat{\beta}_\alpha(i)) = \tilde{0}_{p_\alpha}$ with $\tilde{0}_{p_\alpha}$ a $p_\alpha \times 1$ zero vector. Here

$$U(\beta_\alpha) = D_{\alpha i}^T V_{\alpha i}^{-1}(Y_i - \mu_\alpha^{(i)})/\sigma^2$$

is the quasi-score function. In this expression the components of the matrix $D_{\alpha i}$ of order $(i-1) \times p_\alpha$ are $D_{\alpha i,jk} = \partial \mu_{\alpha j}/\partial \beta_{\alpha k}$, $V_{\alpha i} = diag\{V_1(\mu_{\alpha 1}), \cdots, V_{i-1}(\mu_{\alpha(i-1)})\}$ and the mean vector $\mu_\alpha^{(i)} = (\mu_{\alpha 1}, \cdots, \mu_{\alpha(i-1)})^T$, where the $\mu_{\alpha j}$'s are the proposed means of Y_j's under $\mathcal{M}_\alpha$

Starting with an arbitrary initial value $\hat{\beta}_\alpha^{(0)}(i)$ sufficiently close to $\hat{\beta}_\alpha(i)$, which is supposed to exist, a sequence of parameter estimates generated by the Newton-Raphson method with Fisher scoring is

$$\hat{\beta}_\alpha^{(1)}(i) = \hat{\beta}_\alpha^{(0)}(i) + (\hat{D}_{\alpha i}^{(0)T} \hat{V}_{\alpha i}^{(0)-1} \hat{D}_{\alpha i}^{(0)})^{-1} \hat{D}_{\alpha i}^{(0)T} \hat{V}_{\alpha i}^{(0)-1}(Y_i - \hat{\mu}_\alpha^{(i)(0)}(i)) \qquad (3.2.2)$$

and the quasi-likelihood estimate $\hat{\beta}_\alpha(i)$ may be approached by subsequent iterations.

Instead of starting with an arbitrary initial value, we may also start the iteration with $\hat{\beta}_\alpha^0(i) = \hat{\beta}_\alpha(i-1)$ as long as $\hat{\beta}_\alpha(i-1)$ is available. If $\hat{\beta}_\alpha(i)$ converges as $i \to \infty$, this will be a more efficient way to compute the sequential maximum quasi-likelihood estimates $\{\hat{\beta}_\alpha(i)\}$. Such a technique has been used in Jain (1983).

Having thus obtained $\hat{\beta}_\alpha(i)$, the estimate $\hat{\mu}_{\alpha i}(i)$ of the proposed mean $\mu_{\alpha i}$ of the i-th response value y_i, based on the first $i-1$ observations, can be obtained through (3.2.1). It is in fact the predicted value of the future observation y_i based on the first $i-1$ observations.

Definition: The predictive quasi-deviance function for the vector Y under the generalized linear model (3.2.1) is defined as

$$S_{\alpha,n} = \frac{1}{n}\sum_{i=1}^{n}\int_{\hat{\mu}_{\alpha i}(i)}^{y_i}\frac{y_i - t}{V_i(t)}dt. \tag{3.2.3}$$

Noting that $S_{\alpha,n} = \frac{1}{2n}\sum_{i=1}^{n} D(y_i;\hat{\mu}_{\alpha i}(i))$, where $D(y_i;\hat{\mu}_{\alpha i}(i))$ is the prediction error in terms of the quasi-deviance function for a future value y_i, we have (3.2.3) as the sum of the "honest" prediction errors.

It is apparent that for a model $\mathcal{M}_\alpha$ of the form (3.2.1), the estimated mean values $\hat{\mu}_{\alpha 1}(1), \cdots, \hat{\mu}_{\alpha p_\alpha}(p_\alpha)$ can not be determined since the corresponding $\hat{\beta}_\alpha(i)$ $(i = 1, \cdots, p_\alpha)$ can not be calculated from (3.2.2). The large sample behavior of (3.2.3) are not affected by these first $p_\alpha \leq p$ mean values, so we can set arbitrary finite values to $\hat{\mu}_c\ (i)$ where $i \leq p_\alpha \leq p$. However the arbitrary setting of these mean values do affect the finite sample performance of (3.2.3). One possible way to reduce this effect is to rearrange the order of the first p_α response value Y_{p_α} as follows. First set $\hat{\mu}_{\alpha 1}(1) = 0$ or some other prescribed value. Then choose as the first data point $y_{(1)}$ the one from Y_{p_α} which can be predicted best, i.e. the one with the smallest $D(y_{(1)};\hat{\mu}_{\alpha 1}(1))$. As $y_{(2)}$ we select the nearest data point among Y_{p_α} to $y_{(1)}$ in terms of the quasi-deviance $D(y_{(2)};y_{(1)})$ and define $\hat{\mu}_{\alpha 2}(2) = y_{(1)}$. Then fit $y_{(1)}$ and $y_{(2)}$ using the generalized linear model containing only the first parameter of β_α, calculate the prediction value for each of Y_{p_α} except $y_{(1)}$ and $y_{(2)}$ based on this model, and choose as $y_{(3)}$ the observation among Y_{p_α} which gives the smallest quasi-deviance. The corresponding prediction value for $y_{(3)}$ is defined as $\hat{\mu}_{\alpha 3}(3)$. The next step is to fit the generalized linear model containing the first two parameters of β_α, find $y_{(4)}$ and define $\hat{\mu}_{\alpha 4}(4)$. Continue this procedure until the new order of Y_{p_α} is determined and $\hat{\mu}_{\alpha 1}(1), \cdots, \hat{\mu}_{\alpha p_\alpha}(p_\alpha)$ are defined. While the procedure just discussed can determine the first p_α estimated means and control the prediction errors of the first p_α terms of (3.2.3) to some extent, it still has some disadvantages. One is that a different model will probably yield a different ordering of the data. The model comparisons will thus not be based on the same order of the data. The other is the amount of computation

needed to determine the order which could be fairly large. In Section 3.4 we will again discuss the strategy to deal with the effect of the first p_α estimated mean values.

If the y_i's are normally distributed with constant variance, (3.2.3) becomes the sum of the squared "honest" prediction errors, which was shown by Rissanen (1989) to be an approximation of the predictive stochastic complexity of Y except for a multiplicative constant. If the y_i's have a likelihood function of the form (3.1.2), it can be seen by a straightforward calculation that $D(y_i;\mu_i)$ is the same as the log-likelihood function of y_i with respect to μ_i, except for a term which does not involve the parameter of interest μ_i (part of Theorem 2 in Wedderburn, 1974). Thus in this case (3.2.3) is (except for a quantity unrelated to the model selection) the predictive stochastic complexity of Y relative to model (3.2.1), or equivalently the negative prequential log-likelihood of the model on Y. In general, the use of (3.2.3) as an empirical assessment of the model is an extension of Rissanen's predictive minimum description length principle and Dawid's prequential statistical approach.

3.3 The Predictive Least Quasi-deviance Criterion

3.3.1 Main Results

Because of the above connections between (3.2.3) and the predictive stochastic complexity on the one hand, and prequential analysis on the other, a model selection procedure similar to those employed in the predictive minimum description length principle and "prequential" principle can be formulated for the generalized linear model selection: from the class of models $\mathcal{A}$ we select the one which minimizes $S_{\alpha,n}$, or the most parsimonious one if it is not unique. This is the so-called predictive least quasi-deviance (PLQD) criterion.

We know that a generalized linear model depends not only on the covariate (predictor) variables, but also on the link function $g(\cdot)$. For a fixed $g(\cdot)$ we can use the

predictive least quasi-deviance criterion to select a model $\mathcal{M}_g$ with the smallest $S_{\alpha,n}$ among the models of the form (3.2.1). If another link function $g_1(\cdot)$ is proposed, the same procedure can be applied to obtain another model $\mathcal{M}_{g_1}$. The predictive quasi-deviance function (3.2.3) for these two models can then be compared and the smaller one is selected. The same procedure can be used in the case of more than two link functions. The problem of comparing two different model classes has recently been discussed by O'Hagan (1994).

So far we have considered only the model selection from a set of models $\mathcal{A}$ corresponding to a fixed link function $g(\cdot)$. Now we consider the question of optimality of the model selection by the predictive least quasi-deviance criterion.

Denote $i_{\beta_\alpha} = D_{\alpha i}^T V_{\alpha i}^{-1} D_{\alpha i}$, and $j_{\beta_\alpha} = D_{\alpha i}^T V_{\alpha i}^{-1} V_{\omega i} V_{\alpha i}^{-1} D_{\alpha i}$ where $\omega = \{1, \cdots, p\}$ indicating the full model. Also denote $I_{\beta_\alpha} = -\sigma^2 \frac{\partial U(\beta_\alpha)}{\partial \beta_\alpha}$ where $U(\beta_\alpha)$ is the quasi-score function based on the first $i-1$ observations and the model $\mathcal{M}_\alpha$ of the form (3.2.1). With these notations we have the following results.

Theorem 3.3.1 *Suppose the components of the response n-vector Y are independent with mean $\mu = (\mu_1, \cdots, \mu_n)^T$ and each with a variance $\sigma^2 V_i(\mu_i)$ where σ^2 is a constant of probably unknown value and $V_i(\cdot)$ is a known functions. We consider selecting a model $\mathcal{M}_\alpha$ of the form (3.2.1) from $\mathcal{A}$, given the matrix of observed covariate vectors $X = (x_1, \cdots, x_n)^T$. Suppose furthermore that the following conditions are satisfied.*

(a).

$$E\left|\int_{\hat{\mu}_{\alpha i(i)}}^{\mu_i} \frac{1}{V_i(t)} dt\right| < \infty \quad \text{for any } i \geq 1. \tag{3.3.1}$$

(b).

$$\sum_{i=1}^{\infty} \frac{\sigma^2 V_i(\mu_i)}{i^2} E\left[\int_{\hat{\mu}_{\alpha i(i)}}^{\mu_i} \frac{1}{V_i(t)} dt\right]^2 < \infty. \tag{3.3.2}$$

(c).

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{E\hat{\mu}_{\alpha i(i)}}^{\mu_i} \frac{\mu_i - t}{V_i(t)} dt > 0 \quad \text{for any } \mathcal{M}_\alpha \text{ in Category I.} \tag{3.3.3}$$

(d). $i_{\beta_\alpha} = O(i)$ and $j_{\beta_\alpha} = O(i)$ for all β_α. Moreover, both $i^{-1} i_{\beta_\alpha}$ and j_{β_α} have positive definite limit as $i \to \infty$.

(e). $I_{\beta_\alpha} = O(i)$ *and* I_{β_α} *is nonsingular almost surely as* $i \to \infty$.

(f). $E\left(U(E\hat{\beta}_\alpha(i))\right) = O(i^{1/2})$.

(g). The maximum quasi-likelihood estimate $\hat{\beta}_\alpha(i)$ *of* β_α *in (3.2.1) exists for all* i *greater than* $p_\alpha + 1$.

(h). The link function $g(\cdot)$ *is second-order differential and* g^{-1} *is well defined.*

(i). There exists a positive constant δ *such that* $V(\cdot) > \delta$.

Then

$$S_{\alpha,n} = \frac{1}{2n}D(Y;\mu) + \frac{1}{n}\sum_{i=1}^{n}\int_{E\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{\mu_i - t}{V_i(t)}dt + o_p(1) \qquad (3.3.4)$$

if $\mathcal{M}_\alpha$ *is in Category I, and*

$$S_{\alpha,n} = \frac{1}{2n}D(Y;\mu) + o_p(1) \qquad (3.3.5)$$

if $\mathcal{M}_\alpha$ *is in Category II.*

Furthermore

$$\lim_{n\to\infty} pr\{\textit{the selected model is in Category } I\} = 0. \qquad (3.3.6)$$

Equation (3.3.6) follows directly from (3.3.4), (3.3.5) and (3.3.3) and answers the question of whether the selected model is asymptotically correct. However it may not be optimal. The following theorem gives a result concerning selecting the optimal model.

Theorem 3.3.2 *In addition to the assumptions of Theorem 3.3.1, suppose that the following conditions are true.*

(j). $V_i(\mu_i)$ *is second-order differentiable for* $i = 1, \cdots, n$.

(k). For any model $\mathcal{M}_\alpha$ *in Category II,* $|\hat{\mu}_{\alpha i}(i) - \mu_i| \leq o(1)$ *almost surely as* $i \to \infty$.

(l). For any $\mathcal{M}_\alpha$ *in Category II,* $I_{\beta'_\alpha} - i_{\beta'_\alpha} = o(i)$ *holds for any* β'_α *in* $o(1)$ *neighborhood of true* β_α.

Then for any $\mathcal{M}_\alpha$ in Category II

$$E(S_{\alpha,n}) = \frac{1}{2n}E(D(Y;\mu)) + \frac{1}{2n}\sum_{i=1}^{n}\frac{1}{V_i(\mu_i)}E(\hat{\mu}_{\alpha i}(i) - \mu_i)^2 + o(n^{-1}\log n) \quad (3.3.7)$$

and there exists a non-negative number c_n such that for n sufficiently large

$$E(S_{\alpha,n} - S_{\alpha^*,n}) \geq c_n n^{-1}\log n + o(n^{-1}\log n). \quad (3.3.8)$$

Here $\alpha^* \subset \alpha$ corresponds to any model in Category II which is nested to $\mathcal{M}_\alpha$. Certainly the optimal model is nested to $\mathcal{M}_\alpha$.

The proofs of Theorem 3.3.1 and 3.3.2 are given in Section 3.6.

In Dawid (1992) the consistency problem of the Bayesian model selection by the prequential approach was also considered. It was shown that the model-selection method which proceeds by maximizing the adjusted prequential likelihood, or equivalently minimizing the "adjusted stochastic complexity" of the data, would be (almost surely) consistent.

3.3.2 Remarks on Some Conditions of Theorem 3.3.1 and 3.3.2

Note that if the likelihood function of the data is of the form (3.1.2) and $g(\cdot)$ is a canonical link function, then $D_{\alpha i} = V_{\alpha i}X_\alpha^{(i)}$ where $X_\alpha^{(i)} = (x_{1\alpha},\cdots,x_{(i-1)\alpha})^T$ being an $(i-1)\times p_\alpha$ matrix. Therefore, $\imath_{\beta_\alpha} = X_\alpha^{(i)T}V_{\alpha i}X_\alpha^{(i)}$, $j_{\beta_\alpha} = X_\alpha^{(i)T}V_{\omega i}X_\alpha^{(i)}$ and $I_{\beta_\alpha} = i_{\beta_\alpha}$. Hence the conditions (d), (e) and (l) are obviously true in this situation if $X = O(1)$.

We know that $\sum_1^\infty \frac{1}{i\log^{1+\epsilon}i} < \infty$ for any $\epsilon > 0$. So if $V_i(\mu_i)E\left[\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{1}{V_i(t)}dt\right]^2$ is bounded by $O(i/\log^{1+\epsilon}i)$, the condition (b) will hold. When the likelihood function is of the form (3.1.2) and $g(\cdot)$ is a canonical link function

$$\begin{aligned}E\left[\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{1}{V_i(t)}dt\right]^2 &= E\left[x_{i\alpha}^T\beta_\alpha - x_{i\alpha}^T\hat{\beta}_\alpha(i)\right]^2 \\ &= x_{i\alpha}^T cov(\hat{\beta}_\alpha(i))x_{i\alpha} + x_{i\alpha}^T(E\hat{\beta}_\alpha(i) - \beta_\alpha)(E\hat{\beta}_\alpha(i) - \beta_\alpha)^T x_{i\alpha} = O(1)\end{aligned}$$

using Lemma 3.6.1. Similarly condition (a) also holds in this case.

If n_δ denotes the number of terms satisfying $\int_{E\hat{\mu}_{\alpha_i}(i)}^{\mu_i} \frac{\mu_i - t}{V_i(t)} dt > \delta$ for $i = 1, \cdots, n$, then a sufficient condition implying (c) is

$$\liminf_{n\to\infty} \frac{n_\delta}{n} > 0 \quad \text{for some } \delta > 0.$$

Obviously n_δ/n is an empirical probability for a function of the covariate vector x.

An explanation for the use of condition (k) in (3.3.2) is that the convergence in probability does not imply the convergence in expectation unless the integrand function in the expectation is dominated almost surely by a integral function.

3.4 An Approximate PLQD and A Monte Carlo PLQD

The predictive least quasi-deviance principle has a great intuitive appeal. For one thing, if there is any mechanism which restricts a future observation in a manner similar to the past, and which can be captured by the selected class of parametric functions, then we will find that mechanism. Conversely, if no such mechanism exists, then our predictions will be bad, but so will all other predictions that use the same class of parametric functions. Moreover, the criterion we seek to minimize expresses the quantity which does not involve the hypothetical "true" distribution itself, namely, the accumulated prediction errors of the observations (or the predictive stochastic complexity of the data if the likelihood function of the data is of the form (3.1.2).). Finally, the principle involves a few arbitrary choices that need to be made by "sound judgment"; such choices are the selection of the parametric class and the link functions which, however, are inevitable.

A drawback of the predictive least quasi-deviance technique is that the prediction errors for the first few response observations may be fairly large, and if the sample size n is not large enough the predictive quasi-deviance function (3.2.3) may be seriously affected by these large prediction errors.

To overcome this difficulty, we drop the first few terms of (3.2.3) which do not

affect the asymptotic behavior, but which are always troublesome when computing the predictive quasi-deviance values. The number of terms dropped is proportional to the number of the explanatory variables available. We call this the approximate predictive least quasi-deviance method (abbreviated as APLQD). According to our experience, the number of terms dropped is about γp, where γ is a finite positive number less than 5, so that large prediction errors are avoided and little information is lost.

As another modification to the predictive least quasi-deviance method we consider the following resampling technique: Draw (without replacement) a random collection $\mathcal{C}$ of r permutations of $\{1, \cdots, n\}$ and select a model by minimizing

$$S_{\alpha,n}^{MPLQD} = \frac{1}{r} \sum_{c \in \mathcal{C}} \tilde{S}_{\alpha,n}^{c}$$

where $\tilde{S}_{\alpha,n}^{c}$ is the predictive quasi-deviance value computed by using the approximate predictive least quasi-deviance method based on the permutation $c \in \mathcal{C}$ and $r = O(n)$. This is called the Monte Carlo predictive least quasi-deviance method (abbreviated as MPLQD). Because of the resampling technique used here, the Monte Carlo predictive least quasi-deviance method greatly reduces the effect of large initial prediction errors which seriously affects the performance of the predictive least quasi-deviance method, and is expected to have higher efficiency than the approximate predictive least quasi-deviance method, especially for medium sample size. This can be interpreted as follows. In using the approximate predictive least quasi-deviance method the effect of large initial prediction errors is reduced by dropping the first few terms of (3.2.3), but at the same time the information originating from predicting these deleted observations by other observations is lost. The Monte Carlo predictive least quasi-deviance method seems to compensate for the lost information by resampling. Asymptotically the Monte Carlo predictive least quasi-deviance method behaves similarly to the predictive least quasi-deviance method: (3.3.6) as well as (3.3.8) hold, with the probability being interpreted as the joint probability corresponding to Y and the Monte Carlo selection of the permutations.

Besides the above proposals to solve the effect of the first few terms of (3.2.3), this

initialization problem, inherent in the general predictive principle, can be overcome as follows. The data sequence is subdivided into segments of length d, and each subsequent segment is predicted with the model fitted to the preceding segments. If the very first segment is predicted in the same manner by all the models, the segment length d can be optimized along with the numbers of parameters in the models. Such a technique has been applied successfully to neural networks (Rissanen (1994)). Use of such a technique can also be found in density estimation (Yu and Speed, 1992).

3.5 A Simulation Study

In this section we assess the finite sample performance of the predictive least quasi-deviance method. For the purpose of comparison with other methods we first consider the linear model selection and choose the following example from Shao(1993).

Example 3.5.1: Consider the following model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i,$$

where $i = 1, \cdots, 40$, e_i's are identically independently distributed with the standard normal distribution $N(0,1)$, x_{ki} is the ith value of the kth predictor variable x_k, $x_{1i} = 1$, and the value of x_{ki}, $k = 2, \cdots, 5$, $i = 1, \cdots, 40$, listed in Table 1, are taken from Gunst and Mason(1980). Some of the β_k's may be zero. Thus we are selecting some predictor variables from five possible variables $\{x_1, \cdots, x_5\}$ and we wish to select a model with the best predictive ability. Note that there are thirty-one possible models, and each model is denoted by a subset of $\{1, \cdots, 5\}$ which contains the indices of the variables x_k in the model.

Because y_i's are normally distributed with constant variance, the quasi-deviance function of y_i's is the usual quadratic function. We consider the two modified predictive least quasi-deviance methods: the APLQD and the MPLQD given in Section 4.2 with $\gamma = 2$ (the first 9 terms are dropped when using the approximate predictive least quasi-deviance method and the Monte Carlo predictive least quasi-deviance method)

and $r = 80(= 2n)$. Then we compare the results with the ones obtained by cross-validation methods: the Monte Carlo cross-validation (MCCV(n_v)) and the approximate cross-validation (APCV(n_v)) with $n_v = 25$ and $b = 2n$ (for details about the Monte Carlo cross-validation and the approximate cross-validation techniques and related simulation results see Shao,1993). Table 2 and Table 3 give the empirical probabilities(based on 1000 repetitions) of selecting each model in several different cases.

In this last section we assess the finite sample performance of the predictive least quasi-deviance method. For the purpose of comparison with other methods we choose the following example from Shao(1993).

The following is a summary of the results in Table 2 and Table 3.

1. In terms of the probability of selecting the optimal model, the Monte Carlo predictive least quasi-deviance method and the Monte Carlo cross-validation have the best overall performance among the four methods considered.

2. When the true model has fewer parameters the Monte Carlo cross-validation is slightly better than the Monte Carlo predictive least quasi-deviance technique. However, for the full model the Monte Carlo cross-validation is the worst among all the criteria and the Monte Carlo predictive least quasi-deviance method is the best one.

3. The probability of selecting a model from Category I (incorrect model) is negligible for all four methods if all $\beta_i\bar{x}_i$ values are quite comparable to σ, the standard deviation of the error. Here $\bar{x}_i$ is the sample mean of x_i. If, however, some of the $\beta_i\bar{x}_i$ values are relatively small comparing to σ, this probability can not be controlled.

4. Although the approximate predictive least quasi-deviance method selects the optimal model in expectation, its performance is not as good as expected. This indicates that in order to have a better performance the approximate predictive least quasi-deviance method may require a larger sample size.

In addition to the good performance of the Monte Carlo predictive least quasi-deviance method for linear model selection, it provides a powerful tool for generalized linear model selection as well. The following two examples illustrates the performance of the predictive least quasi-deviance method to generalized linear model selection.

Example 3.5.2: Consider the following generalized linear model

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

in which the response y has a Poisson distribution. The observations of the predictor variables x's are given in Table 4. The sample size is 36. The values of y are generated from the Poisson distribution with the mean given by the above model. Thus we can obtain the empirical probability of selecting the optimal model by the two modified predictive least quasi-deviance methods. The results at $\gamma = 4.75$ (or drop the first 19 terms when compute the predictive quasi-deviance) and $r = 10, 20, 40, 80$ are listed in Table 5. From Table 5 we find the Monte Carlo predictive least quasi-deviance method generally has a quite satisfying performance. It also implies that the approximate predictive least quasi-deviance method would require a larger size sample to achieve a better performance. In addition, the probability of selecting a model from Category I is negligible for the Monte Carlo predictive least quasi-deviance method if all $\beta_i \bar{x}_i$ values are comparable to each other and are not close to zero; it is not the case otherwise.

Example 3.5.3: The data in Table 6, taken from Schoener(1970), have already been analyzed by Fienberg(1970), Bishop *et al.*(1975) and by McCullagh and Nelder (1989, Section 4.6). Data concerning the daytime habits of two species of lizard, *grahami* and *opalinus*, were collected by observing occupied sites or perches and recording the appropriate description, namely the species involved, the time of day, the height and the diameter of the perch and whether the site was sunny or shaded. The purpose of analyzing this set of data is to compare the two species with regard to their preferred perches.

We now consider an analysis using a linear logistic model fitted by maximum likelihood suggested in McCullagh and Nelder (1989). Since there are four factors H,

D, S and T available, we have a class of models $\mathcal{C}$ each member of which includes some of the four factors and one two-factor interaction. There are 40 different models in this class and the largest model contains 8 independent variables. The response variable here is the observed number of sites occupied by *grahami* lizards or, equivalently, the observed proportion of total sites that were occupied by *grahami* lizards. We then choose a model in $\mathcal{C}$ which fits the observations of the response variable best by using the Monte Carlo predictive least quasi-deviance method with $\gamma \approx 2.5$(or drop the first 19 terms in computing the predictive quasi-deviance) and $r = 50(\approx 2n)$. The resulting optimal model can be written symbolically as $H+D+S+T$, the second best model is $H + D + S + T + H.D$ and the estimated predictive deviance for these two models are 0.0618 and 0.0632 respectively. Based on the optimal model, comparison with regard to the preferred perches of the two species can be obtained.

Our conclusion is the same as that of McCullagh and Nelder (1989), who informally used the analysis of deviance method to remove the interaction term from the model, but different from that of Fienberg (1970) and Bishop *et al.*(1975), who found significant interaction between H and D and between S and T regarding their effect on species' preferences. The reason for the difference appears to be the fact that, as pointed in McCullagh and Nelder (1989), these authors attempted to consider several unrelated issues simultaneously using only a single model, and did not condition on the totals of the occupied sites which are regarded as ancillary in the method of generalized linear regression.

3.6 Proofs of Theorems 3.3.1 and 3.3.2

The following result in martingale theory will be useful in the sequel (see Section 3.3 of Stout,1974).

Let $\{Z_i, \mathcal{F}_i, i \geq 1\}$ be a sequence of martingale differences, where $\mathcal{F}_i$ is the borel field generated by $Z_1, \cdots, Z_i$, $i \geq 1$, $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Define $S_n = \sum_{i=1}^{n} Z_i$.

Theorem 3.6.1 *[Chow, 1960,1967]*

If for some $r \geq 2$

$$\sum_{i=1}^{\infty} E\,|Z_i|^r\,/i^{1+r/2} < \infty,$$

then $S_n/n \to 0$ *almost everywhere as* $n \to \infty$.

Lemma 3.6.1 *Under the conditions (d)-(h) of Theorem 3.3.1, the following statements are true.*

(i). $cov(\hat{\beta}_\alpha(i)) = O(i^{-1})$ *and* $\hat{\beta}_\alpha(i) - E(\hat{\beta}_\alpha(i)) = O_p(i^{-1/2})$.

(ii). $E\left(\hat{\beta}_\alpha(i) - \beta_\alpha\right)\left(\hat{\beta}_\alpha(i) - \beta_\alpha\right)^T = O(i^{-1})$ *and* $\hat{\beta}_\alpha(i) - \beta_\alpha = O_p(i^{-1/2})$ *if* $\mathcal{M}_\alpha$ *is in Category II.*

(iii). $var(\hat{\mu}_{\alpha i}(i)) = O(i^{-1})$ *and* $\hat{\mu}_{\alpha i}(i) - E(\hat{\mu}_{\alpha i}(i)) = O_p(i^{-1/2})$.

(iv). $E\,|\hat{\mu}_{\alpha i}(i) - \mu_i|^2 = O(i^{-1})$ *and* $\hat{\mu}_{\alpha i}(i) - \mu_i = O_p(i^{-1/2})$ *if* $\mathcal{M}_\alpha$ *is in Category II.*

Proof. If we denote $\hat{\beta}_\alpha(i)$ as the maximum quasi-likelihood estimate, then $U(\hat{\beta}_\alpha(i)) = 0$. Applying Taylor expansion to $U(\hat{\beta}_\alpha(i))$ around $E\hat{\beta}_\alpha(i)$ for any $\mathcal{M}_\alpha$ in Category I, it can be seen that

$$\tilde{0}_{p_\alpha} = U(E\hat{\beta}_\alpha(i)) - \sigma^{-2} I_{\beta_\alpha^*}(\hat{\beta}_\alpha(i) - E\hat{\beta}_\alpha(i))$$

where $I_{\beta_\alpha^*}$ is the observed information evaluated at a point β_α^* lying on the line segment joining $\hat{\beta}_\alpha(i)$ and $E\hat{\beta}_\alpha(i)$. Thus

$$\hat{\beta}_\alpha(i) - E\hat{\beta}_\alpha(i) = \sigma^2 I_{\beta_\alpha^*}^{-1} U(E\hat{\beta}_\alpha(i)) = O(i^{-1})U(E\hat{\beta}_\alpha(i)). \tag{3.6.1}$$

By the condition $var(U(E\hat{\beta}_\alpha(i))) = j_{E\hat{\beta}_\alpha(i)} = O(i)$ and the condition (f), we have from (3.6.1) $cov(\hat{\beta}_\alpha(i)) = O(i^{-1})$ and by Chebyshev's inequality $\hat{\beta}_\alpha(i) - E\hat{\beta}_\alpha(i) = O_p(i^{-1/2})$, which is (i).

By expanding $U(\hat{\beta}_\alpha(i))$ around β_α, similarly we can obtain (ii).

Using the expansion

$$\begin{aligned}|\hat{\mu}_{\alpha i}(i) - E\hat{\mu}_{\alpha i}(i)| &= \left|g^{-1}(x_{i\alpha}^T\hat{\beta}_\alpha(i)) - g^{-1}(x_{i\alpha}^T E\hat{\beta}_\alpha(i))\right. \\ &\quad \left. + g^{-1}(x_{i\alpha}^T E\hat{\beta}_\alpha(i)) - Eg^{-1}(x_{i\alpha}^T\hat{\beta}_\alpha(i))\right| \\ &\leq \left|g^{-1}(x_{i\alpha}^T\beta_\alpha^*)\right|\left|x_{i\alpha}^T(\hat{\beta}_\alpha(i) - E\hat{\beta}_\alpha(i))\right| \\ &\quad + E\left[\left|g^{-1}(x_{i\alpha}^T\beta_\alpha^*)\right|\left|x_{i\alpha}^T(\hat{\beta}_\alpha(i) - E\hat{\beta}_\alpha(i))\right|\right]\end{aligned}$$

and part (i), it can be seen (iii) is true. Similarly (iv) follows. □

Lemma 3.6.2 $((A,\ B)^TC(A,\ B))^{-1} - \begin{pmatrix} (A^TCA)^{-1} & 0 \\ 0 & 0 \end{pmatrix}$ *is a non-negative definite matrix, where* C *is an* $m \times m$ *positive definite matrix,* A *is an* $m \times a$ *matrix,* B *is a* $m \times b$ *matrix,* $a + b \leq m$ *and rank*$(A,\ B) = a + b$.

The proof of Lemma 3.6.2 is straightforward.

Proof of Theorem 3.3.1. Using the notation of Section 3.2, (3.2.3) can be rewritten as

$$\begin{aligned}S_{\alpha,n} &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mu_i}^{y_i}\frac{y_i - t}{V_i(t)}dt + \frac{1}{n}\sum_{i=1}^{n}\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{y_i - \mu_i}{V_i(t)}dt \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\int_{E\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{\mu_i - t}{V_i(t)}dt + \frac{1}{n}\sum_{i=1}^{n}\int_{\hat{\mu}_{\alpha i}(i)}^{E\hat{\mu}_{\alpha i}(i)}\frac{\mu_i - t}{V_i(t)}dt \\ &\stackrel{\text{def}}{=} \frac{1}{2n}D(Y;\mu) + I_1 + I_2 + I_3 \qquad (3.6.2)\end{aligned}$$

where I_1, I_2, I_3 denote the second, third and fourth term of the right hand side. This decomposition of $S_{\alpha,n}$ can be explained as follows. The first term is half of the average quasi-deviance of Y, $I_2 + I_3$ measures the bias sequence $\{\mu_i - \hat{\mu}_{\alpha i}(i)\}$, where I_2 measures the $\{\mu_i - E\hat{\mu}_{\alpha i}(i)\}$ part and I_3 for $\{\hat{\mu}_{\alpha i}(i) - E\hat{\mu}_{\alpha i}(i)\}$, and I_1 is some kind of cross-product term.

Since $\hat{\mu}_{\alpha i}(i)$ depends on the first $i - 1$ observations we have

$$E\left(\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{y_i - \mu_i}{V_i(t)}dt \mid y_1, \cdots, y_{i-1}\right) = (E(y_i - \mu_i))\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i}\frac{1}{V_i(t)}dt = 0$$

$i = 2, \cdots, n$. Thus $\{\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i} \frac{y_i - \mu_i}{V_i(t)} dt, i = 1, \cdots, n, \cdots\}$ is a sequence of martingale differences if $E\left|\int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i} \frac{y_i - \mu_i}{V_i(t)} dt\right| < \infty$, which is true from the condition (3.3.1). By Theorem 3.6.1 it is known that $I_1 \to 0$ almost surely as $n \to \infty$ if (3.3.2) is true.

From (iii) of Lemma 3.6.1 and by Lyapunov's inequality it is easy to know that $E\,|\hat{\mu}_{\alpha i}(i) - E(\hat{\mu}_{\alpha i}(i))| = O(i^{-1/2})$. Therefore it can be seen that under the condition (i) I_3 converges to 0 in L_1 and accordingly $I_3 = o_p(1)$ in probability.

Suppose that the chosen model $\mathcal{M}_\alpha$ is the correct one. From (iv) of Lemma 3.6.1 it is easily known that $E\,|\hat{\mu}_{\alpha i}(i) - \mu_i| = O(i^{-1/2})$. Therefore

$$\lim_{i \to \infty} \int_{E\hat{\mu}_{\alpha i}(i)}^{\mu_i} \frac{\mu_i - t}{V_i(t)} dt = 0$$

and $\lim_{n \to \infty} I_2 = 0$ if $\mathcal{M}_\alpha$ is in Category II.

For any model $\mathcal{M}_\alpha$ in Category I and $\mathcal{M}_\gamma$ in Category II it is not difficult to comprehend that the difference of the corresponding predictive quasi-deviance functions

$$S_{\alpha,n} - S_{\gamma,n} > 0 \quad \text{in probability}$$

if (3.3.3) is true. Note that $I_2 > 0$ for any fixed n when $\mathcal{M}_\alpha$ is in Category I. (3.3.3) is therefore a quite reasonable assumption.

Theorem 3.3.1 follows from the above results for I_1, I_2 and I_3 and the decomposition (3.6.2) of $S_{\alpha,n}$. □

Proof of Theorem 3.3.2. Rewrite $S_{\alpha,n}$ as

$$S_{\alpha,n} = \frac{1}{2n} D(Y; \mu) + \frac{1}{n} \sum_{i=1}^{n} \int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i} \frac{y_i - \mu_i}{V_i(t)} dt + \frac{1}{n} \sum_{i=1}^{n} \int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i} \frac{\mu_i - t}{V_i(t)} dt.$$

Define $f(s) = \int_s^{\mu_i} \frac{\mu_i - t}{V_i(t)} dt$ and by Taylor expansion for $f(s)$ around μ_i we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \int_{\hat{\mu}_{\alpha i}(i)}^{\mu_i} \frac{\mu_i - t}{V_i(t)} dt = \frac{1}{2n} \sum_{i=1}^{n} \left[\frac{1}{V_i(\mu_i)} (\hat{\mu}_{\alpha i}(i) - \mu_i)^2 + O((\hat{\mu}_{\alpha i}(i) - \mu_i)^3) \right].$$

Thus by the independence of y_i and $\hat{\mu}_{\alpha i}(i)$, and the condition (k) it is easily known that

$$\begin{aligned} ES_{\alpha,n} &= \frac{1}{2n} ED(Y; \mu) + \frac{1}{2n} \sum_{i=1}^{n} \left[\frac{1}{V_i(\mu_i)} E(\hat{\mu}_{\alpha i}(i) - \mu_i)^2 + o(E(\hat{\mu}_{\alpha i}(i) - \mu_i)^2) \right] \\ &= \frac{1}{2n} ED(Y; \mu) + \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{V_i(\mu_i)} E(\hat{\mu}_{\alpha i}(i) - \mu_i)^2 + o(n^{-1} \log n) \end{aligned}$$

since $E|\hat{\mu}_{\alpha i}(i) - \mu_i| = O(i^{-1/2})$. From (3.6.1) and the condition (l) $\hat{\beta}_\alpha(i) - \beta_\alpha = I^{-1}_{\beta^*_\alpha} U(\beta_\alpha) = (1 + o(1)) i^{-1}_{\beta_\alpha} U(\beta_\alpha)$. Thus by Taylor expansion

$$\begin{aligned}
\hat{\mu}_{\alpha i}(i) - \mu_i &= g^{-1}(x_{i\alpha}^T \hat{\beta}_\alpha(i)) - g^{-1}(x_{i\alpha}^T \beta_\alpha) \\
&= (g^{-1})'(x_{i\alpha}^T \beta_\alpha)(x_{i\alpha}^T \hat{\beta}_\alpha(i) - x_{i\alpha}^T \beta_\alpha) + o(x_{i\alpha}^T \hat{\beta}_\alpha(i) - x_{i\alpha}^T \beta_\alpha) \\
&= (g^{-1})'(x_{i\alpha}^T \beta_\alpha)(1 + o(1))(x_{i\alpha}^T \hat{\beta}_\alpha(i) - x_{i\alpha}^T \beta_\alpha) \\
&= (g^{-1})'(x_i^T \beta)(1 + o(1)) x_{i\alpha}^T i^{-1}_{\beta_\alpha} U(\beta_\alpha).
\end{aligned}$$

Therefore by noting that $V_{\alpha i} = V_{\omega i}$, $\mu_\alpha^{(i)} = \mu_\omega^{(i)}$ and $x_{i\alpha}^T \beta_\alpha = x_i^T \beta$ for any $\mathcal{M}_\alpha$ in Category II we have

$$E(\hat{\mu}_{\alpha i}(i) - \mu_i)^2 = \left[(g^{-1})'(x_i^T \beta)\right]^2 (1 + o(1)) x_{i\alpha}^T i^{-1}_{\beta_\alpha} x_{i\alpha}.$$

Similarly

$$E(\hat{\mu}_{\alpha^* i}(i) - \mu_i)^2 = \left[(g^{-1})'(x_i^T \beta)\right]^2 (1 + o(1)) x_{i\alpha^*}^T i^{-1}_{\beta_{\alpha^*}} x_{i\alpha^*}.$$

From Lemma 3.6.2 we know $i i^{-1}_{\beta_\alpha} - \begin{pmatrix} i i^{-1}_{\beta_{\alpha^*}} & 0 \\ 0 & 0 \end{pmatrix}$ is a non-negative definite matrix, therefore

$$\begin{aligned}
E(S_{\alpha,n} - S_{\alpha^*,n}) &= \frac{1}{2n} \sum_{i=1}^{n} \frac{\left[(g^{-1})'(x_i^T \beta)\right]^2}{i V_i(\mu_i)} x_{i\alpha}^T \left[i i^{-1}_{\beta_\alpha} - \begin{pmatrix} i i^{-1}_{\beta_{\alpha^*}} & 0 \\ 0 & 0 \end{pmatrix} \right] x_{i\alpha} + o(n^{-1} \log n) \\
&\geq c_n n^{-1} \log n + o(n^{-1} \log n) \qquad \text{if } n \text{ is sufficiently large.}
\end{aligned}$$

□

Table 3.1: *The Values of x_{kt} in Example 3.5.1*

x_{2t}	x_{3t}	x_{4t}	x_{5t}	x_{2t}	x_{3t}	x_{4t}	x_{5t}
0.36	0.53	1.06	0.5326	0.09	0.18	0.59	0.1855
1.32	2.52	5.74	3.6183	0.02	0.16	0.24	0.1572
0.06	0.09	0.27	0.2594	0.02	0.11	0.21	0.0998
0.16	0.41	0.83	1.0346	0.05	0.24	0.43	0.2804
0.01	0.02	0.07	0.0381	0.11	0.39	0.29	0.2879
0.02	0.07	0.07	0.3440	0.18	0.11	0.43	0.6810
0.56	0.62	2.12	1.4559	0.04	0.09	0.23	0.3242
0.98	1.06	2.89	4.0182	0.85	1.33	2.70	2.6013
0.32	0.20	0.76	0.4600	0.17	0.32	0.66	0.4469
0.01	0.00	0.07	0.1540	0.08	0.12	0.49	0.2436
0.15	0.25	0.50	0.6516	0.38	0.18	0.49	0.4400
0.24	0.28	0.59	0.0611	0.11	0.13	0.18	0.3351
0.11	0.35	0.40	0.1922	0.39	0.38	0.99	1.3979
0.08	0.13	0.28	0.0931	0.43	0.46	1.47	2.0138
0.61	0.85	0.49	0.0538	0.57	1.16	1.82	1.9356
0.03	0.03	0.23	0.0199	0.13	0.03	0.08	0.1050
0.06	0.11	0.50	0.0419	0.04	0.05	0.14	0.2207
0.02	0.08	0.25	0.1093	0.13	0.18	0.28	0.0180
0.04	0.24	0.08	0.0328	0.20	0.95	0.41	0.1017
0.00	0.02	0.04	0.0797	0.07	0.06	0.18	0.0962

Table 3.2: Probabilities (Based on 1000 Repetitions) of Selecting Each Model

	Model	Category	*MCCV*	*APCV*	*APLQD*	*MPLQD*
$\beta=$	1,4	Optimal	0.926	0.501	0.532	0.921
(2,0,0,4,0)	1,2,4	II	0.023	0.116	0.108	0.025
	1,3,4	II	0.021	0.085	0.069	0.020
	1,4,5	II	0.028	0.172	0.211	0.033
	1,2,3,4	II	0.002	0.038	0.012	0.001
	1,2,4,5	II	0.000	0.039	0.038	0.000
	1,3,4,5	II	0.000	0.037	0.021	0.000
	1,2,3,4,5	II	0.000	0.012	0.009	0.000
$\beta=$	1,4,5	Optimal	0.956	0.651	0.785	0.951
(2,0,0,4,8)	1,3,5	I	0.001	0.000	0.000	0.000
	1,2,4,5	II	0.026	0.161	0.113	0.038
	1,3,4,5	II	0.016	0.131	0.078	0.011
	1,2,3,4,5	II	0.001	0.057	0.024	0.000
$\beta=$	1,4,5	I	0.019	0.000	0.002	0.011
(2,9,0,4,8)	1,2,4,5	Optimal	0.956	0.818	0.797	0.962
	1,3,4,5	I	0.003	0.000	0.000	0.007
	1,2,3,4,5	II	0.022	0.182	0.201	0.020
$\beta=$	1,2	I	0.000	0.002	0.000	0.000
(2,9,0,4,0.1)	1,4	I	0.010	0.017	0.018	0.005
	1,2,3	I	0.000	0.003	0.000	0.000
	1,2,4	I	0.949	0.496	0.773	0.910
	1,2,5	I	0.000	0.003	0.000	0.000
	1,3,4	I	0.001	0.017	0.001	0.000
	1,3,5	I	0.000	0.002	0.000	0.000
	1,4,5	I	0.000	0.019	0.000	0.000
	1,2,3,4	I	0.020	0.199	0.107	0.024
	1,2,3,5	I	0.000	0.005	0.000	0.000
	1,2,4,5	Optimal	0.020	0.158	0.086	0.054
	1,3,4,5	I	0.000	0.016	0.002	0.000
	1,2,3,4,5	II	0.000	0.063	0.013	0.007

Table 3.3: Continued to Table 3.2

	Model	Category	*MCCV*	*APCV*	*APLQD*	*MPLQD*
$\beta=$	1,2,3,5	I	0.001	0.000	0.000	0.001
(2,9,6,4,8)	1,2,4,5	I	0.006	0.000	0.002	0.001
	1,3,4,5	I	0.031	0.001	0.005	0.023
	1,2,3,4,5	Optimal	0.962	0.999	0.993	0.975
$\beta=$	1	I	0.000	0.002	0.000	0.000
(2,0.3,0,0,1.4)	1,2	I	0.045	0.042	0.018	0.057
	1,3	I	0.000	0.006	0.001	0.000
	1,4	I	0.019	0.033	0.021	0.018
	1,5	I	0.900	0.418	0.632	0.886
	1,2,3	I	0.001	0.011	0.003	0.001
	1,2,4	I	0.000	0.007	0.000	0.000
	1,2,5	Optimal	0.025	0.104	0.030	0.028
	1,3,4	I	0.000	0.016	0.016	0.000
	1,3,5	I	0.008	0.118	0.179	0.007
	1,4,5	I	0.000	0.078	0.049	0.002
	1,2,3,4	I	0.000	0.013	0.006	0.001
	1,2,3,5	II	0.002	0.044	0.014	0.000
	1,2,4,5	II	0.000	0.042	0.010	0.000
	1,3,4,5	I	0.000	0.043	0.020	0.000
	1,2,3,4,5	II	0.000	0.023	0.001	0.000

Table 3.4: *The Values of* x_{ki} *in Example 3.5.2*

x_{1i}	x_{2i}	x_{3i}	x_{1i}	x_{2i}	x_{3i}	x_{1i}	x_{2i}	x_{3i}
0.412	0.284	0.97	0.484	0.885	0.83	0.993	0.51	0.469
0.805	0.296	1.082	0.249	0.969	0.108	0.784	0.577	0.404
0.485	0.23	0.743	0.443	0.949	0.363	0.754	0.559	1.031
0.235	0.173	1.038	0.594	0.948	1.195	0.964	0.53	0.742
0.224	0.16	0.796	0.541	0.959	1.147	0.729	0.502	0.711
0.31	0.136	0.832	0.464	1.003	0.564	1.13	0.559	0.92
0.262	0.285	0.387	1.15	1.045	0.906	0.896	0.515	1.169
0.51	0.103	0.436	0.982	0.948	0.221	0.672	0.54	1.032
0.614	0.208	0.88	0.734	1.042	0.711	1.484	0.51	0.358
0.453	0.204	0.224	0.628	0.924	0.833	0.931	0.533	0.355
0.095	0.197	0.67	0.944	0.985	0.731	0.566	0.5	0.653
0.841	0.533	0.531	0.562	0.938	0.36	0.841	0.533	0.531

Table 3.5: Probabilities (Based on 1000 Repetitions) of Selecting Each Model

	Model	Category	*APLQD*	*MPLQD*			
				$r = 10$	$r = 20$	$r = 40$	$r = 80$
$\beta =$	1	I	0.008	0.000	0.000	0.000	0.000
(2.1 [illegible])	1,2	Optimal	0.571	0.693	0.711	0.710	0.719
	1,3	I	0.003	0.000	0.000	0.000	0.000
	1,4	I	0.002	0.000	0.000	0.000	0.000
	1,2,3	II	0.182	0.150	0.138	0.136	0.131
	1,2,4	II	0.182	0.132	0.117	0.123	0.125
	1,2,3,4	II	0.052	0.025	0.034	0.031	0.025
$\beta =$	1,2,3	Optimal	0.777	0.842	0.856	0.860	0.861
(2,2,2,0)	1,2,3,4	II	0.223	0.158	0.144	0 140	0.139
$\beta =$	1,2	I	0.015	0.001	0.000	0.000	0.000
(2,1,0.5,0.35)	1,3	I	0.001	0.000	0.000	0.000	0.000
	1,2,3	I	0.144	0.066	0.061	0.056	0.060
	1,2,4	I	0.040	0.004	0.005	0.006	0.003
	1,2,3,4	Optimal	0.800	0.929	0.934	0.938	0.937
$\beta =$	1,2	I	0.350	0.403	0.400	0.412	0.413
(2,3,0,0.1)	1,2,3	I	0.111	0.083	0.080	0.081	0.081
	1,2,4	Optimal	0.417	0.426	0.423	0.423	0.429
	1,2,3,4	II	0.122	0.088	0.097	0.084	0.077

Table 3.6: A comparison of site preferences of two species of lizards, *grahami* and *opalinus*

			T								
	Perch		*Early*			*Mid-day*			*Late*		
S	*D* (in)	*H* (ft)	*G*	*O*	*Total*	*G*	*O*	*Total*	*G*	*O*	*Total*
Sun	≤2	<5	20	2	22	8	1	9	4	4	8
		≥5	13	0	13	8	0	8	12	0	12
	>2	<5	8	3	11	4	1	5	5	3	8
		≥5	6	0	6	0	0	0	i	1	2
Shade	≤2	<5	34	11	45	69	20	89	18	10	28
		≥5	31	5	36	55	4	59	13	3	16
	>2	<5	17	15	32	60	32	92	8	8	16
		≥5	12	1	13	21	5	26	4	4	8

H, perch height; *D*, perch diameter; *S*, sunny/shady; *T*, time of day; *G*, *grahami*; *O*, *opalinus*.

Chapter 4

On Stochastic Complexity Estimation — A Decision Theoretic Approach

4.1 Introduction

The raw material of a statistical investigation is a set of observations, which are the observed values of some random variable X whose distribution F is at least partly unknown. Statistical inference is concerned with methods of using this observational material to obtain information concerning the probabilistic structure of F. A general formulation of the problem was given by Wald's theory of decision procedures (Section 1.1 of Wald, 1950 and Chapter 1 of Ferguson, 1967) according to which the aim of statistics is the selection of a decision rule which minimizes the resulting risk.

For the purpose of describing the information contained in the observational material the related notions of stochastic complexity and description length provide a global measure in the sense that the derivation of these quantities involves the consideration of not only the randomness in the observational material but also the properties of mathematical formulation used to model the observations (refer to Chapter 1). With such measures one can hope to determine a decision procedure with some

universal optimum properties for many statistical problems.

Let $X_1, X_2, \cdots, X_n = X^n$ be a sample independently drawn from a (at least partly unknown) probability density function $p(\cdot)$ which is assumed to belong to a density class Γ. The X_i's are assumed to take values in a measurable space $\mathcal{X}$ and the density function $p(\cdot)$ is taken with respect to a known complete, σ-finite dominating measure $\nu(\cdot)$. The description length for the sample X^n relative to p is then defined as a two-step codelength

$$C(p) + \log \frac{1}{p(X^n)}, \tag{4.1.1}$$

where $p(X^n) = \prod_{i=1}^n p(X_i)$, $C(p)$ is the part of the code length for encoding the underlying density p and the logarithm is in base 2.

An interpretation as well as some necessary restrictions are given as follows. Suppose the class Γ contains at most countable infinite number of densities, $\{C(p), p \in \Gamma\}$ is then a sequence of nonnegative numbers satisfying Kraft's inequality $\sum_{p\in\Gamma} 2^{-C(p)} \leq 1$ and each $C(p)$ is interpreted to be the codelength for the description of the corresponding density. There is also a Bayesian interpretation of the numbers $2^{-C(p)}$ as prior probabilities. In Kolmogorov's complexity theory $C(p)$ is equal to the minimum codelength of the programs φ that encode p on a universal computer which consists of finite length binary programs satisfying the prefix property (Rissanen, 1989 pp. 45-52). Since the description of $X_1, X_2, \cdots, X_n$ in (4.1.1) follows the code for p, the prefix condition is essential for decoding the two steps. Kraft's inequality gives necessary and sufficient conditions for the prefix property, i.e. the existence of instant and decodable binary codes. By Shannon's work, if p is given, then $\lceil \log(1/p(X^n)) \rceil$ is the length of an instantaneous code that describes the sample $X_1, X_2, \cdots, X_n$.

In order to make Kraft's inequality meaningful the countability of the size of the density class Γ is necessary. Although in statistical inference the employed model class is often of uncountable size (like the usual parametric model class), we can circumvent the problem by applying the encoding process to the parameters which are truncated to a fixed precision and then, by using a limit process, to extend it to a model class of uncountable size. The countability of the model class is, therefore,

of no importance to the results following. Nevertheless, for the sake of simplicity we will assume the countability of Γ, except in Section 4.3. Hence Γ is often specified by a sequence of parametric models with the parameter values restricted to a prescribed precision, and in the ideal case Γ consists of all computable probability densities.

When the descriptive programs for the densities in the class Γ are determined, the two-step codelength (4.1.1) is a function of the unknown part of the true density p_0 provided that p_0 is in the class Γ. It is natural to consider the minimization of the two-step length to determine the unknown part of p_0. This is the so-called minimum description length (MDL) method. However, the minimum description length based on (4.1.1) is still not entirely satisfactory as the shortest codelength of the sample $X_1, X_2, \cdots, X_n$ relative to the density class Γ, for it is the result of a specific coding construct, and by encoding both the sample and the density we get more than we really need. To eliminate such redundancy in describing the data, a concept of stochastic complexity is introduced by Rissanen (1986a, 1987, 1989). In our case, the stochastic complexity of the sample $X_1, X_2, \cdots, X_n$ relative to Γ and C is defined as

$$I(X^n \mid \Gamma, C) = -\log\left(\sum_{p\in\Gamma} p(X^n)2^{-C(p)}\right). \tag{4.1.2}$$

It represents the shortest code length for the data that can be achieved by the densities in Γ under C. Based on the stochastic complexity a density can be defined to replace the true density in the inferential process.

Now suppose that the unknown part of the density p can be written as $\phi = \phi(p)$ which is a mapping, called the descriptive mapping, from a large class of densities $\mathcal{P}$ to a space $\mathcal{A}$, called the description space. The large class $\mathcal{P}$ contains the plausible densities p for the unknown population density of X^n, which usually includes Γ, the convex hull of Γ Γ^*, all the empirical densities of X^n, and so on. In the most general situation $\mathcal{P}$ contains all the probability density functions. The structure of $\mathcal{A}$ and the form of ϕ are determined by the particular decision problem. Any function (or mapping) $\delta = \delta(X^n)$ that maps the sample space $\mathcal{X}^n$ into $\mathcal{A}$ is called a *decision rule*. The class of all decision rules is denoted by D, the decision space.

In a decision problem, if p is fixed, $\phi(p)$ will be completely determined. A general decision problem, in which p is at least partly unknown, is then specified by assigning p a correct decision in the decision space D to estimate $\phi(p)$ using the information contained in the observations and using a loss function L to evaluate such decision. The great variety of the possible decision structures is illustrated by the following cases:

i) *Hypothesis testing* in which one wishes to decide which of the propositions A_1 or A_2 is true for the density p_0. Here $\mathcal{A} = \{A_1, A_2\}$ and $\phi(p_0) = A_1$ if A_1 is true and A_2 if A_2 is true. The decision rule $\delta(X^n)$ can be any function taking the value of either 1 or 0 corresponding to A_1 and A_2 respectively.

ii) *Identification* A straightforward generalization of (i) in which there is a choice of s alternatives $A_1, A_2, \cdots, A_s$. Here $\mathcal{A} = \{A_1, A_2, \cdots, A_s\}$ and $\phi(p_0) = A_i$ if A_i is true, $i = 1, 2, \cdots, s$.

iii) *Estimation* on the other hand requires a numerical assessment of some quantity related to the unknown part of p_0. In this case $\phi = \phi(p)$ is a k-vector functional of p and $\mathcal{A}$ is the Euclidean space $\mathcal{R}^k$ or its subset.

For an account of these cases in somewhat different forms see Rissanen (Chapter 4, 1989) for (i), Rissanen and Ristad (1992) and Rissanen (Chapter 7, 1989) for (ii), and Barron and Cover (1991), Barron et al. (1992), and Rissanen et al. (1992) for (iii). In this chapter we address these problems within the framework of decision theory. The choice of a loss function is still to be discussed.

For a decision rule $\delta = \delta(X^n)$, we define the loss functions, respectively, to each case above as follows.

i)

$$L(\phi(p), A_1) = \begin{cases} 0 & \text{if } \phi(p) = A_1, \\ a & \text{if } \phi(p) = A_2, \end{cases} \tag{4.1.3}$$

$$L(\phi(p), A_2) = \begin{cases} 0 & \text{if } \phi(p) = A_2, \\ b & \text{if } \phi(p) = A_1, \end{cases} \tag{4.1.4}$$

a and b are the losses which can be adjusted according to the relative importance of the two types of error.

The risk function becomes

$$\begin{aligned} R(\phi(p),\delta) &= E_p L(\phi(p),\delta(X^n)) \\ &= \begin{cases} bP(\delta(X^n)=A_2) & \text{if } \phi(p)=A_1 \\ aP(\delta(X^n)=A_1) & \text{if } \phi(p)=A_2 \end{cases} \end{aligned} \tag{4.1.5}$$

ii)

$$L(\phi(p),A_j) = r_{ij} \tag{4.1.6}$$

for $\phi(p)=A_i, j=1,2,\cdots,s$ and $i=1,2,\cdots,s$. r_{ij} is the penalty for misclassifying $\phi(p)$ of proposition A_i to proposition A_j. $r_{ij}=0$ if $i=j$.

The risk function is

$$R(\phi(p),\delta) = \sum_{j=1}^{s} r_{ij} P(\delta(X^n)=A_j) \tag{4.1.7}$$

for $\phi(p)=A_i, i=1,2,\cdots,s$.

iii)

$$L(\phi(p),\delta(X^n)) = v(p)|\delta(X^n)-\phi(p)|^2, \tag{4.1.8}$$

the usual form of the squared error loss.

From now on we use the triplet (Γ, D, R) to denote a statistical decision problem.

In this chapter we concentrate on estimating $\phi(p)$ of the density p. By using Rissanen's concept of stochastic complexity we introduce a complexity decision rule: first we define a stochastic complexity density estimate $\tilde{p}_n$ of p with respect to the sample $X_1, X_2, \cdots, X_n$, the density class Γ and the description length sequence $C(p)$, then we use $\phi(\tilde{p}_n)$ to estimate the quantity $\phi(p)$. We show that this decision procedure is admissible, achieves the minimum expected risk and forms a minimal complete class under very general conditions. Applications to parametric distribution families is also considered after a discussion of consistency.

4.2 Complexity Decision Rule

To estimate the quantity $\phi(p)$, a general and straightforward selection procedure can be obtained from a set of candidate probability densities Γ subject to the information provided by the sample $X_1, X_2, \cdots, X_n$. Denote the density chosen as $\hat{p}_n$, the estimate of $\phi(p)$ can then be written as $\phi(\hat{p}_n)$ and the decision rule $\delta(X^n) = \phi(\hat{p}_n)$.

The concept of description length and stochastic complexity suggests that a natural and optimal choice for the density p might be the one that minimizes the description length or the one generated by the minimum stochastic complexity. We now discuss such procedure in detail.

Let's first define *minimum description length* of the data $X_1, X_2, \cdots, X_n$ relative to Γ and C as

$$B(X^n) = \min_{p\in\Gamma}(C(p) + \log\frac{1}{p(X^n)}) \tag{4.2.1}$$

and

$$\hat{p}_n = \arg\min_{p\in\Gamma}(C(p) + \log\frac{1}{p(X^n)}) \tag{4.2.2}$$

which is called the *minimum description length density estimator* relative to Γ and C (Barron and Cover, 1991). In case of ties, the density $\hat{p}_n$ is chosen for which $C(\hat{p}_n)$ is shortest (and any further ties are broken by selecting the density with the smallest index in Γ).

From Section 4.1 we know that $\{C(p), p \in \Gamma\}$ must satisfy the summability requirement

$$\sum_{p\in\Gamma} 2^{-C(p)} \leq 1. \tag{4.2.3}$$

In the remainder of this chapter we assume, for the purpose of convenience, that $\{C(p), p \in \Gamma\}$ satisfies the *regularity condition*

$$\sum_{p\in\Gamma} 2^{-C(p)} = 1. \tag{4.2.4}$$

Actually, if the regularity condition is not satisfied, we can define a new coding process for the densities in Γ so that the description length for each p in Γ is the $C'(p)$

that satisfies

$$2^{-C'(p)} = \frac{2^{-C(p)}}{\sum_{q\in\Gamma} 2^{-C(q)}} \quad \text{or} \quad C'(p) = C(p) + \log\left(\sum_{q\in\Gamma} 2^{-C(q)}\right). \tag{4.2.5}$$

Thus for the new sequence $\{C'(p), p \in \Gamma\}$ the regularity condition (4.2.4) holds and the minimum description length of X^n relative to Γ and C' differs from the one relative to Γ and C by only the constant $\log(\sum_{q\in\Gamma} 2^{-C(q)})$.

Now for each density p in the candidate class Γ and for the corresponding description length $C(p)$ there exists a coding process in which the length of the codeword for each sample $X^n = X_1, X_2, \cdots, X_n$, written as $b(X^n \mid p, C)$, is equal to

$$C(p) + \log \frac{1}{p(X^n)} \tag{4.2.6}$$

in difference of a constant less than 1. The corresponding binary code is instantaneous and decodable so that Kraft's inequality holds. This coding process can be described by the following coding system.

Suppose each observation X_i is observed to a prescribed precision α. The measurable space $\mathcal{X}$ can then be quantized into a countable alphabet $[\mathcal{X}]$ over which the observation ranges. Let us write $[\mathcal{X}]^n$ for the set of all observation strings of length n and $[\mathcal{X}]^* = \bigcup_{n=0}^{\infty} [\mathcal{X}]^n$ for their union. Let B denote the binary alphabet and $B^* = \bigcup_{n=0}^{\infty} B^n$. We define a *coding system* relative to Γ and C as a (decoding) function

$$G : \Omega \to [\mathcal{X}]^* \tag{4.2.7}$$

from a subset Ω of B^* onto $[\mathcal{X}]^*$. Here Ω is the set of all codewords of observation strings obtained by the coding process defined by Γ and C. Any member b_i of Ω, such that $G(b_i) = X^n$ is said to be a codeword of the sample X^n. The length $|b_i|$ of b_i is the number of binary digits in it. It is easy to see that the inverse image of X^n under the decoding map G is $G^{-1}(X^n) = \{b(X^n \mid p, C), p \in \Gamma\}$. By the regularity condition (4.2.4) the sum

$$p'(X^n) = \sum_{p\in\Gamma} 2^{-|b(X^n|p,C)|} = \sum_{p\in\Gamma} p(X^n) 2^{-C(p)} \tag{4.2.8}$$

is well defined (indeed, its integral on the measurable space $\mathcal{X}^n$ equals to one). In fact, $p'(X^n)\nu(\Delta[X^n])$ represents the probability of finding a codeword for X^n in a game of fair coin tossings, where $\Delta[X^n]$ denotes the quantization region containing X^n. Hence we get the *stochastic complexity* of $X^n = X_1, X_2, \cdots, X_n$ relative to Γ and C as

$$I(X^n \mid \Gamma, C) = -\log\left(\sum_{p\in\Gamma} p(X^n)2^{-C(p)}\right). \tag{4.2.9}$$

It can be regarded as the code length obtained by the removal of the redundancy in the coding system G and represents the shortest code length for the data X^n that can be achieved by the densities in Γ under C (Rissanen, 1989, pp. 45-67).

By the criterion (4.2.2) a minimum description length (MDL) density estimator $\hat{p}_n(X)$ can be obtained for the observations $X_1, X_2, \cdots, X_n$, which exists with probability one (Barron and Cover, 1991). For a future observation X a natural question that may be asked is if the MDL density $\hat{p}_n$ would still produce the minimum description length for $X^nX = X_1, X_2, \cdots, X_n, X$. The answer is negative because the MDL density estimator depends on the observed sample. The following example provides an illustration.

Example 4.2.1 Let $X_1, X_2, \cdots, X_n$ be a sample from a normal distribution $N(j, 1)$ where j takes some positive integer $1, 2, \cdots$. The density description length $C(N(j, 1))$ $= j$, so that regularity condition (4.2.4) is satisfied.

For each $N(j, 1)$ the description length for the sample X^n is

$$j + \frac{1}{2}\left(\sum_{i=1}^{n}(X_i - j)^2\right)\log e + n\log\sqrt{2\pi}. \tag{4.2.10}$$

Expression (4.2.10) is minimized when $j = \hat{j}_n = [m_1(X^n) - 1/(n\log e) + 1/2]$ where $m_1(X^n) = (1/n)\sum_{i=1}^{n} X_i$ is the sample mean. The MDL density estimator $N(\hat{j}_n, 1)$ changes according to the sample. However, when n is sufficiently large, by the law of large numbers $\hat{j}_n$ is uniquely determined and so is the MDL density estimator.

In general we have the following result obtained in Barron and Cover (1991).

Theorem 4.2.1 *(Barron and Cover, 1991) Suppose that $C(p)$ satisfies the regularity condition (4.2.4). If the true density $p_0 \in \Gamma$, then*

$$\hat{p}_n \equiv p_0 \tag{4.2.11}$$

for all sufficiently large n, with probability one.

The following corollary for the estimator $\phi(\hat{p}_n)$ holds.

Corollary 4.2.1 *Suppose that $\phi(p)$ can be expressed in the form of a linear functional*

$$\phi(p) = E_p f(X) = \int f(x)p(x)\nu(dx) \tag{4.2.12}$$

for any $p \in \Gamma$, where $f(x)$ is a ν-measurable function with $E_p|f(x)| < \infty$ and is called a kernel of ϕ. If $p_0 \in \Gamma$, then

$$\phi(\hat{p}_n) \equiv \phi(p_0) \tag{4.2.13}$$

for all sufficiently large n, with probability one.

A disadvantage of the MDL estimator is that it yields no closed expression for the estimator until a concrete form of p as well as Γ is given. In addition, because of the redundancy in the two-stage coding system, the MDL (4.2.1) overestimates the real code length.

Motivated by such considerations, we define a density, called the *stochastic complexity density estimator* as

$$\tilde{p}_n(x) \stackrel{\text{def}}{=} \tilde{p}_n(x|X^n) = \frac{\sum_{p\in\Gamma} p(x)p(X^n)2^{-C(p)}}{\sum_{p\in\Gamma} p(X^n)2^{-C(p)}} \tag{4.2.14}$$

which is generated by the difference between the stochastic complexity of X^nX and the stochastic complexity of X^n, i.e. by the stochastic complexity of X given X^n.

To see this notice that $-\log\left(\sum_{p\in\Gamma} p(X)p(X^n)2^{-C(p)}\right)$ is the stochastic complexity for X^nX and $-\log\left(\sum_{p\in\Gamma} p(X^n)2^{-C(p)}\right)$ is the stochastic complexity for X^n. From the fact that $\tilde{p}_n$ has integral one with respect to x on the measurable space $\mathcal{X}$ and by (4.2.8) we have the following proposition.

Proposition 4.2.1 *For almost every sample X^n $\tilde{p}_n$ is a probability density that satisfies (4.2.14) for almost every $x \in \mathcal{X}$.*

We have obtained a new estimate $\phi(\tilde{p}_n)$ of $\phi(p)$ which we call the *stochastic complexity estimator.* Note that $\tilde{p}_n$, a reasonable estimate of the density for X^n, may not belong to Γ since Γ is not necessarily a convex set. This is of no consequence because Γ is only a proposed density class, and the assumption that the true density belongs to Γ may not be true. Further, the quantity of interest is $\phi(p)$ rather than the density p itself, so even if the true density p_0 is in Γ, $\phi(q)$ may still be the same as $\phi(p_0)$ for some q outside Γ. The relationship between $\phi(\tilde{p}_n)$ and the minimum description length estimator $\phi(\hat{p}_n)$ is established by the following result.

Theorem 4.2.2 *Suppose that $C(p)$ satisfies the regularity condition (4.2.4) and that for any given sample X^n $\hat{p}_n$ is the MDL density estimator defined by (4.2.2). Suppose, moreover, that for any $p \in \Gamma$*

$$C(p) + \log \frac{1}{p(X^n)} = C(\hat{p}_n) + \log \frac{1}{\hat{p}_n(X^n)} + \tau^2(p, X^n) \tag{4.2.15}$$

where $\hat{p}_n(X^n) = \prod_{i=1}^n \hat{p}_n(X_i)$ and $\tau^2(p, X^n)$ is a positive functional satisfying

$$\lim_{n\to\infty} (\tau^2(p, X^{n+1}) - \tau^2(p, X^n)) = 0 \tag{4.2.16}$$

uniformly for $p \in \Gamma$. Then for each x, except a set N of measure 0,

$$\lim_{n\to\infty} \tilde{p}_n(x) = \lim_{n\to\infty} \hat{p}_n(x) = p_0(x) \tag{4.2.17}$$

with probability one, where $\tilde{p}_n$ is defined by (4.2.14) and $p_0 \in \Gamma$ is the true density of X^n.

Proof: We know from Section 4.1 that X is a random variable defined on the probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P)$. By Kolmogorov's theorem on the extension of measures (Shiryayev, 1984, pp. 161), there exists a probability space $(\mathcal{X}^\infty, \mathcal{B}^\infty(\mathcal{X}), P)$ for the sequence $X_1, X_2, \cdots, X_n, \cdots$. Now from Theorem 4.2.1 there exists a $\mathcal{B}^\infty(\mathcal{X})$-measurable set M with $P(M) = 0$ so that for each observation sequence $(X_1, X_2, \cdots, X_n, \cdots) \in M^c$

$$\hat{p}_n \equiv p_0 \quad \text{for all sufficiently large n} \tag{4.2.18}$$

If for each $X \in \mathcal{X}$ we regard $X^n X$ as a new sample, then by (4.2.15)

$$p(X)p(X^n)2^{-C(p)} = \hat{p}_{n+1}(X)\hat{p}_{n+1}(X^n)2^{-C(\hat{p}_{n+1})}2^{-\tau^2(p,X^n,X)} \tag{4.2.19}$$

where $\hat{p}_{n+1} = \hat{p}_{n+1}(\cdot \mid X^n, X)$ is the MDL density estimator based on $X^n X$. (Notice that $\hat{p}_n$ does not depend on the order of X^n, thus the MDL density estimator based on XX^n is also $\hat{p}_{n+1}$.)

Next we show that there exists a $\mathcal{B}(\mathcal{X})$-measurable set N with $P(N) = 0$ so that for each $X \in N^c$

$$\hat{p}_{n+1} \equiv p_0 \quad \text{for all sufficiently large } n \text{ and all } (X_1, \cdots, X_n, \cdots) \in M^c. \tag{4.2.20}$$

Let A be the set of all $X \in \mathcal{X}$ that does not satisfy (4.2.20). Clearly

$$A = \{X \in \mathcal{X} \mid (X, X_1, \cdots, X_n, \cdots) \in M \text{ and } (X_1, \cdots, X_n, \cdots) \in M^c\} \tag{4.2.21}$$

and $A \times M^c \subseteq M$.

Hence $P(A) = 0$ because $0 = P(M) \geq P(A \times M^c) = P(A)P(M^c) = P(A)$, and we obtain $N = A$.

Now for any $x \in N^c$ rewrite $\tilde{p}_n$ as

$$\tilde{p}_n = \frac{\hat{p}_{n+1}(x)\hat{p}_{n+1}(X^n)2^{-C(\hat{p}_{n+1})}\sum_{p\in\Gamma} 2^{-\tau^2(p,x,X^n)}}{\hat{p}_n(X^n)2^{-C(\hat{p}_n)}\sum_{p\in\Gamma} 2^{-\tau^2(p,X^n)}}. \tag{4.2.22}$$

From (4.2.18),(4.2.20),(4.2.4) and (4.2.16) the result (4.2.17) follows. □

A similar consistency result derived from a prequential analysis approach is given by Dawid (1992).

Example 4.2.2 (Example 4.2.1 continued) The stochastic complexity for X^n is

$$\begin{aligned} I(X^n \mid N(j,1), C) &= -\log\left(\sum_{j=1}^{\infty} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - j)^2\right) 2^{-j}\right) \\ &= n\log\sqrt{2\pi} + \frac{1}{2}nm_2(X^n)\log e - \frac{1}{2}nm_1^2(X^n)\log e \\ &\quad -\log\left(\sum_{j=1}^{\infty} \exp\left(-\frac{1}{2}n(j - m_1(X^n))^2\right) 2^{-j}\right) \end{aligned} \tag{4.2.23}$$

where $m_2(X^n) = (1/n)\sum_{i=1}^n X_i^2$. By the concavity of $\log x$,

$$\begin{aligned} -\frac{1}{2}nc\log e &\geq \log\left(\sum_{j=1}^{\infty}\exp\left(-\frac{1}{2}n(j-m_1(X^n))^2\right)2^{-j}\right) \\ &\geq \left(-\frac{1}{2}n\log e\right)\sum_{j=1}^{\infty}(j-m_1(X^n))^2 2^{-j} \\ &= -\frac{1}{2}n\left(6-4m_1(X^n)+m_1^2(X^n)\right)\log e \end{aligned} \tag{4.2.24}$$

where $c > 0$ is a constant. Thus we can obtain an estimated range for $I(X^n \mid N(j,1), C)$

$$\begin{aligned} &\log\sqrt{2\pi} + (m_2(X^n) - m_1^2(X^n))\log\sqrt{e} + c\log\sqrt{e} \\ \leq \frac{1}{n}I(X^n \mid N(j,1), C) &\leq \log\sqrt{2\pi} + (m_2(X^n) - 4m_1(X^n))\log\sqrt{e} + 3. \end{aligned} \tag{4.2.25}$$

The stochastic complexity density estimator $\tilde{p}_n$ for given X^n is

$$\begin{aligned} \tilde{p}_n(x) &= \frac{\sum_{j=1}^{\infty}\frac{1}{(2\pi)^{(n+1)/2}}\exp\left(-\frac{1}{2}\left(\sum_{i=1}^n(X_i-j)^2+(x-j)^2\right)\right)2^{-j}}{\sum_{j=1}^{\infty}\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^n(X_i-j)^2\right)2^{-j}} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2(1-\frac{1}{n+1})+\frac{nm_1(X^n)}{n+1}x-\frac{nm_1^2(X^n)}{2(n+1)}\right)\times \\ &\quad \frac{\sum_{j=1}^{\infty}\exp\left(-\frac{1}{2}(n+1)\left(j-\frac{nm_1(X^n)+x}{n+1}\right)^2\right)2^{-j}}{\sum_{j=1}^{\infty}\exp\left(-\frac{1}{2}n(j-m_1(X^n))^2\right)2^{-j}}. \end{aligned} \tag{4.2.26}$$

Suppose the true density is $N(\mu, 1)$, where μ is some positive integer. Since $E|X_1|^r < \infty$ for any $1 \leq r < 2$, by Marcinkiewicz's strong law of large numbers (Stout,1974, pp. 126)

$$\frac{\sum_{i=1}^n X_i - n\mu}{n^{1/r}} \to 0 \quad \text{a.s..} \tag{4.2.27}$$

Hence

$$m_1(X^n) = \mu + \varepsilon_n \quad \text{a.s.} \tag{4.2.28}$$

where

$$\varepsilon_n = o(n^{-(1-1/r)}) \quad \text{for } 1 \leq r < 2. \tag{4.2.29}$$

Now we can write

$$\frac{\sum_{j=1}^{\infty}\exp\left(-\frac{1}{2}(n+1)\left(j-\frac{nm_1(X^n)+x}{n+1}\right)^2\right)2^{-j}}{\sum_{j=1}^{\infty}\exp\left(-\frac{1}{2}n(j-m_1(X^n))^2\right)2^{-j}}=\frac{w_{n+1}+o(w_{n+1})}{w_n+o(w_n)}\quad\text{a.s.}\qquad(4.2.30)$$

where $w_n=\exp(-nc_n^2/2)2^{-\mu}$. Then it is easy to see that

$$\begin{aligned}\lim_{n\to\infty}\frac{w_{n+1}}{w_n}&=\lim_{n\to\infty}\exp\left(-\frac{1}{2}(n+1)\varepsilon_{n+1}^2+\frac{1}{2}n\varepsilon_n^2\right)\\&=\lim_{n\to\infty}\exp\left(-\frac{1}{2}n(\varepsilon_{n+1}^2-\varepsilon_n^2)-\frac{1}{2}\varepsilon_{n+1}^2\right)=1\quad\text{a.s.}\qquad(4.2.31)\end{aligned}$$

by (4.2.29) and $\varepsilon_{n+1}-\varepsilon_n=-m_1(X^n)/(n+1)+x/(n+1)$. From these results we have

$$\lim_{n\to\infty}\tilde{p}_n(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\mu)^2\right)\quad\text{a.s..}\qquad(4.2.32)$$

4.3 Application to Parametric Families

In this section we study the stochastic complexity estimation in an important class of densities, namely, the parametric families, either with a prior density $\pi(\theta)$ for the paramcters

$$\Gamma_b=\{p(x|\theta),\pi(\theta)\}\qquad(4.3.1)$$

or without one, i.e.

$$\Gamma_e=\{p(x|\theta)\}.\qquad(4.3.2)$$

In both cases $\theta=(\theta_1,\theta_2,\cdots,\theta_k)$ denotes a k-component free parameter, i.e. a vector ranging over a subset Θ of the k-dimensional Euclidean space $\mathcal{R}^k$ with non-empty interior. (Often in such models some of the natural parameters are not free but a relationship, either implicit or explicit exists among them. However, we assume that the dependent parameters have been eliminated and only the free ones remain in the model.)

First we find the stochastic complexity for the sample X^n relative to Γ_b. Notice that there are an uncountable number of densities in Γ_b and thus we cannot construct

a prefix code for each density in it. But by the argument in Rissanen (1989, pp. 53-67), the parameters are considered to be truncated to some finite precision, say θ_j to the precision $\alpha_j = 2^{-q_j}$ where q_j is the number of fractional binary digits taken in the truncation. Then a prefix code can be constructed which assigns to each such truncated parameter vector θ a codeword with length $C(\theta)$ given by the least integral upper bound to $-\log \pi(\theta) - \sum_{j=1}^{k} \log \alpha_j$. The two-step codelength for the sample X^n relative to each truncated parameter is

$$-\log p(X^n|\bar{\theta}) - \log \pi(\bar{\theta}) - \sum_{j=1}^{k} \log \alpha_j \tag{4.3.3}$$

where $p(X^n|\bar{\theta}) = \prod_{i=1}^{n} p(X_i|\bar{\theta})$ and $\bar{\theta}$ denotes the truncated parameter vector θ to the precision $(\alpha_1, \cdots, \alpha_k)$. Letting $\alpha_j \to 0$, $j = 1, 2, \cdots, k$, by (4.2.9) the stochastic complexity of X_n relative to Γ_b goes to the integral

$$I(X_n \mid \Gamma_b) = -\log \left(\int_{\Theta} p(X^n|\theta)\pi(\theta)d\theta \right) \tag{4.3.4}$$

Therefore, with the same argument that yielded (4.2.14) we can define the stochastic complexity density estimator for X_n drawn from Γ_b as

$$\tilde{p}_n(x) = \frac{\int_{\Theta} p(x|\theta)p(X^n|\theta)\pi(\theta)d\theta}{\int_{\Theta} p(X^n|\theta)\pi(\theta)d\theta} \tag{4.3.5}$$

and the stochastic complexity estimator for $\phi(p)$ is $\phi(\tilde{p}_n)$. (4.3.5) as follows. Assume a distribution

For a sample coming from the set of densities $\Gamma_e = \{p(x|\theta)\}$ we can still define a coding system in which a prefix code is constructed for the data for each parameter value. The stochastic complexity is obtained by applying a procedure similar to (4.2.8) and (4.2.9). The key point here is to use a universal prior for integers to define an optimum code for each truncated parameter as well as its enclosing optimum precision. The procedure is to quantize the parameter space with quantization regions of identically shaped rectangles. The shape of this rectangle is decided by the ellipsoid $(\theta - \hat{\theta}_n)^T M(\hat{\theta}_n)(\theta - \hat{\theta}_n) \leq d$ such a way that each rectangle is identical with the maximum intersecting rectangle of this ellipsoid. Here $\hat{\theta}_n$ is the maximum likelihood

estimate, $M(\theta)$ is the Hessian matrix of the double derivatives of $-\log p(x|\theta)$ and d is any fixed number which will be optimized further. These quantization regions are ordered according to their natural distance $\|\theta\|_{M(\theta)} = \sqrt{\theta^T M(\theta)\theta}$ whereby a sequence of integers are obtained indicating their position. Then we use the universal prior to provide a codelength for each position-index integer.

The universal prior assigns each integer a codelength

$$C^*(n-1) = \log c_0 + \log^* n, \quad n = 1, 2, \cdots \tag{4.3.6}$$

where c_0 is a constant of about 2.865064 and $\log^* y = \log y + \log\log y + \cdots$, where only the positive terms are included in the sum. This length function $C^*(n)$ has the optimum property that for any distribution $P(n)$ for the positive integers such that

i) $P(n) \geq P(n+1), n > M$ for some M

ii) $-\sum_{n\geq 1} P(n)\log P(n) = \infty$,

the following holds

$$\lim_{N\to\infty} \frac{\sum_{n=0}^{N} P(n)C^*(n)}{-\sum_{n=0}^{N} P(n)\log P(n)} = 1 \tag{4.3.7}$$

which indicates that we could do no better even if a distribution $P(n)$ to design the code with were given.

The universal prior $2^{-C^*(n)}$ can be considered a modification of the improper prior $\{1/n\}$ of Jeffreys which is sometimes used to express complete ignorance. It is derived from coding of integers in a manner that certain natural coding theoretic requirements are satisfied and it objectively expresses one's initial ignorance when this notion is made precise (see Elias (1975) and Rissanen (1983) for details). When some constraints for the parameters exist the universal prior for the integers, which presupposes no prior knowledge, should no longer be used in the coding process. In that case Jaynes' maximum entropy principle could guide us to construct a prior distribution for the two-step codelength of the sample (Jaynes,1978).

By the derivation of Rissanen (1983) the two-step codelength for X^n relative to Γ_e is

$$-\log p(X^n|\theta) + \log^* \left(V(k)(\|\theta\|_{M_n(\theta)})^k\right) \tag{4.3.8}$$

where $M_n(\theta)$ is the Hessian matrix of the double derivatives of function $-\log p(X^n|\theta)$ and $V(k)$ is the volume of the k-dimensional unit ball

$$V(k) = \begin{cases} (2\pi)^{k/2} / (k/2)!2^{k/2} & k \text{ even}, \\ \pi^{(k-1)/2}2^{k+1}((k+1)/2)!/(k+1)! & k \text{ odd}. \end{cases} \tag{4.3.9}$$

The stochastic complexity of X^n relative to Γ_e is

$$I(X^n \mid \Gamma_e) = -\log \int_\Theta p(X^n|\theta) c_1 2^{-\log^*(V(k)(\|\theta\|_{M_n(\theta)})^k)} d\theta \tag{4.3.10}$$

where c_1 is a constant satisfying

$$c_1^{-1} = \int_\Theta 2^{-\log^*(V(k)(\|\theta\|_{M_n(\theta)})^k)} d\theta,$$

and the stochastic complexity density estimator for X^n drawn from Γ_e is

$$\tilde{p}_n(x) = \frac{\int_\Theta p(x|\theta) p(X^n|\theta) 2^{-\log^*(V(k)(\|\theta\|_{M_{n+1}(\theta)})^k)} d\theta}{\int_\Theta p(X^n|\theta) 2^{-\log^*(V(k)(\|\theta\|_{M_n(\theta)})^k)} d\theta} \tag{4.3.11}$$

All the quantities (4.3.8), (4.3.10) and (4.3.11) appear to be more of theoretical rather than practical value. When $-\log p(X^n|\theta)$ grows proportionally to n, as normally is the case, the elements of $M_n(\theta)$ are of order n and $\log^*(V(k)(\|\theta\|_{M_n(\theta)})^k)$ is dominated by $\log(V(k)(\|\theta\|_{M_n(\theta)})^k)$. (4.3.8), (4.3.10) and (4.3.11) can then be approximated by substituting $\log(\cdot)$ for $\log^*(\cdot)$.

4.4 Minimum Expected Risk and Admissibility

In the previous sections we proposed a stochastic complexity density estimator and a related stochastic complexity estimator for $\phi(p)$.

To evaluate the quality of the stochastic complexity density estimator we define a *loss functional* $L(p, q)$ for any two densities p and q defined on s-dimensional Euclidean space

$$L(p,q) = \int_{\mathcal{R}^s} |\psi_p(t) - \psi_q(t)|^2 \nu(dt) = \int_{\mathcal{R}^s} (\psi_p(t) - \psi_q(t))(\bar{\psi}_p(t) - \bar{\psi}_q(t)) \nu(dt) \tag{4.4.1}$$

where ψ_p and ψ_q are the characteristic functions of p and q respectively, and $\bar{\psi}$ is the conjugate function of ψ. The characteristic function of p is defined as $\psi_p(t) = \int_{\mathcal{R}^s} p(X)\exp\{i(X,t)\}\nu(dX)$, where (X,t) is the inner product in $\mathcal{R}^s$ and $t \in \mathcal{R}^s$.

The integral $\int_{\mathcal{R}^s} |\psi_p(t)|^2\nu(dt)$ may not always exist and the condition of absolute quadratic integrability of the characteristic function should be assumed here. However, for most of the usual densities this condition is satisfied. There are other possible definitions of the loss functional such as

$$L_1(p,q) = \int_{\mathcal{X}} |p-q|\nu(dX) \tag{4.4.2}$$

which is just the Hellinger distance H_1 and

$$L_2(p,q) = \int_{\mathcal{X}} (p-q)^2\nu(dX) \tag{4.4.3}$$

in which $\int_{\mathcal{X}} p^2\nu(dX) < \infty$ is assumed. (See Devroye, 1987, pp. 1-11).

Notice that when $\mathcal{X}$ is the usual Euclidean space $\mathcal{R}^s$ or its subset the two loss functionals L and L_2 are equivalent in the sense that

$$\int_{\mathcal{R}^s} |\psi_p(t) - \psi_q(t)|^2\nu(dt) = (2\pi)^s \int_{\mathcal{R}^s} (p(X) - q(X))^2\nu(dX),$$

which can be obtained by Parseval-Plancherel formula (Hazewinkel et al, 1991, pp. 163) due to the fact that $\psi_p - \psi_q$ is the Fourier transformation of $p-q$.

Let $X_1, X_2, \cdots, X_n$ be a sample independently drawn from a density p which is assumed to belong to Γ with countable number of densities. The X_i's are assumed to take values in a measurable space $\mathcal{X} \subseteq \mathcal{R}^s$ and the density p is taken with respect to a known complete σ-finite dominating measure ν. $\{C(p), p \in \Gamma\}$ is a sequence of description numbers for the densities in Γ satisfying regularity condition (4.2.4). For any density estimator δ based on the sample X^n the risk functional is

$$R(p,\delta) = E_p L(p,\delta) = \int_{\mathcal{X}^n} \left(\int_{\mathcal{R}^s} |\psi_p(t) - \psi_\delta(t)|^2\nu(dt) \right) p(X^n)\nu(dX^n) \tag{4.4.4}$$

for $p \in \Gamma$. Thus the expected risk is

$$\begin{aligned} r(\delta \mid C) &= \sum_{p\in\Gamma} R(p,\delta)2^{-C(p)} \\ &= \sum_{p\in\Gamma} \int_{\mathcal{X}^n} \left(\int_{\mathcal{R}^s} |\psi_p(t) - \psi_\delta(t)|^2\nu(dt) \right) p(X^n)2^{-C(p)}\nu(dX^n). \end{aligned} \tag{4.4.5}$$

We have the following results.

Theorem 4.4.1 *The stochastic complexity density estimator $\tilde{p}_n$ minimizes the expected risk functional $r(\delta \mid C)$ among all the density estimators of p.*

If there exists some δ' such that $r(\delta' \mid C) < \infty$, then $\tilde{p}_n$ is the unique density estimator minimizing $r(\delta \mid C)$.

Proof: We prove only the second part, the first part easily follows. Let $\delta = \delta(X_1, \cdots, X_n)$ be an arbitrary density estimator of $p \in \Gamma$ based on the sample X^n. Then

$$\begin{aligned} r(\delta \mid C) &= \sum_{p\in\Gamma} \int_{\mathcal{X}^n} \left(\int_{\mathcal{R}^s} |\psi_p(t) - \psi_\delta(t)|^2 \nu(dt) \right) p(X^n) 2^{-C(p)} \nu(dX^n) \\ &= \sum_{p\in\Gamma} \int_{\mathcal{X}^n} \left(\int_{\mathcal{R}^s} |\psi_p(t) - \psi_\delta(t)|^2 \nu(dt) \right) \frac{p(X^n)2^{-C(p)}}{\sum_{q\in\Gamma} q(X^n)2^{-C(q)}} \times \\ &\quad \left(\sum_{q\in\Gamma} q(X^n) 2^{-C(q)} \right) \nu(dX^n) \\ &= \int_{\mathcal{X}^n} \int_{\mathcal{R}^s} \left[\sum_{p\in\Gamma} |\psi_p(t) - \psi_\delta(t)|^2 \frac{p(X^n)2^{-C(p)}}{\sum_{q\in\Gamma} q(X^n)2^{-C(q)}} \right] \times \\ &\quad \left(\sum_{q\in\Gamma} q(X^n) 2^{-C(q)} \right) \nu(dt)\, \nu(dX^n) \end{aligned} \tag{4.4.6}$$

The operations of integral and summation are interchangeable because the integrand is non-negative. $p(X^n)2^{-C(p)} / \left(\sum_{q\in\Gamma} q(X^n)2^{-C(q)} \right)$ can be regarded as a posterior probability density given the observations X^n.

To minimize $r(\delta \mid C)$ is now equivalent of minimizing

$$\sum_{p\in\Gamma} |\psi_p(t) - \psi_\delta(t)|^2 \frac{p(X^n)2^{-C(p)}}{\sum_{q\in\Gamma} q(X^n)2^{-C(q)}} \quad \text{for any fixed } t \text{ and } X^n \tag{4.4.7}$$

It is easy to see that when

$$\psi_\delta(t) = \frac{\sum_{p\in\Gamma} \psi_p(t) p(X^n) 2^{-C(p)}}{\sum_{q\in\Gamma} q(X^n)2^{-C(q)}} = \psi_{\tilde{p}_n}(t) \tag{4.4.8}$$

(4.4.7) is minimized, and for any δ' such that $r(\delta' \mid C) < \infty$, the minimum is the same at least for some t and X^n. Consequently

$$\delta = \tilde{p}_n(x) = \frac{\sum_{p\in\Gamma} p(x) p(X^n) 2^{-C(p)}}{\sum_{q\in\Gamma} q(X^n)2^{-C(q)}} \tag{4.4.9}$$

is the unique density to minimize $r(\delta \mid C)$. □

The same conclusion holds for the loss functional L_2.

Theorem 4.4.2 *The stochastic complexity density estimator $\tilde{p}_n$ minimizes the expected risk*

$$r_2(\delta \mid C) = \sum_{p \in \Gamma} \int_{\mathcal{X}^n} \left(\int_{\mathcal{X}} (\delta - p)^2 \nu(dX) \right) p(X^n) 2^{-C(p)} \nu(dX^n) \tag{4.4.10}$$

among all the density estimators of p.

If there exists some δ' such that $r_2(\delta' \mid C) < \infty$, then $\tilde{p}_n$ is the unique density estimator minimizing $r_2(\delta \mid C)$.

In practice many quantities of interest $\phi(p)$ can be expressed in the form of a linear functional

$$\phi(p) = E_p f(x) = \int_{\mathcal{X}} f(x) p(x) \nu(dx) \tag{4.4.11}$$

If we use the loss function (4.1.8) with $v(p) = 1$ to evaluate the stochastic complexity estimate $\phi(\tilde{p}_n)$, then similarly to Theorem 4.4.1 we get

Theorem 4.4.3 *The stochastic complexity estimator $\phi(\tilde{p}_n)$ minimizes the expected risk function*

$$r_3(\phi, \delta) = \sum_{p \in \Gamma} \int_{\mathcal{X}^n} |\delta - \phi(p)|^2 p(X^n) 2^{-C(p)} \nu(dX^n) \tag{4.4.12}$$

among all the estimators of $\phi(p)$.

If there exists some δ' such that $r_3(\phi, \delta') < \infty$, then $\phi(\tilde{p}_n)$ is the unique estimator minimizing $r_3(\phi, \delta)$.

Remark: Generalization of Theorems 4.4.1,4.4.2 and 4.4.3 to a parametric family follows naturally.

Next we show the admissibility of the stochastic complexity estimator of a density and $\phi(p)$ in the countable set Γ among estimators based on the data $X_1, X_2, \cdots, X_n$. By definition, a density estimator $\hat{p}_n^{(1)}$ (or an estimator $\delta^{(1)}(X^n)$ of $\phi(p)$) is ***inadmissible***

if there is another density estimator $\hat{p}_n^{(2)}$ (or $\delta^{(2)}(X^n)$) such that

$$\begin{aligned}&\int_{\mathcal{X}^n}\left(\int_{\mathcal{R}^s}|\psi_{\hat{p}_n^{(2)}}(t)-\psi_p(t)|^2\nu(dt)\right)p(X^n)\nu(dX^n)\\ \le&\int_{\mathcal{X}^n}\left(\int_{\mathcal{R}^s}|\psi_{\hat{p}_n^{(1)}}(t)-\psi_p(t)|^2\nu(dt)\right)p(X^n)\nu(dX^n)\quad\text{for all } p\in\Gamma\end{aligned}\tag{4.4.13}$$

and

$$\begin{aligned}&\int_{\mathcal{X}^n}\left(\int_{\mathcal{R}^s}|\psi_{\hat{p}_n^{(2)}}(t)-\psi_p(t)|^2\nu(dt)\right)p(X^n)\nu(dX^n)\\ <&\int_{\mathcal{X}^n}\left(\int_{\mathcal{R}^s}|\psi_{\hat{p}_n^{(1)}}(t)-\psi_p(t)|^2\nu(dt)\right)p(X^n)\nu(dX^n)\quad\text{for some } p\in\Gamma\end{aligned}\tag{4.4.14}$$

(or

$$\begin{aligned}&\int_{\mathcal{X}^n}|\delta^{(2)}(X^n)-\phi(p)|^2p(X^n)\nu(dX^n)\\ \le&\int_{\mathcal{X}^n}|\delta^{(1)}(X^n)-\phi(p)|^2p(X^n)\nu(dX^n)\quad\text{for all } p\in\Gamma\end{aligned}\tag{4.4.15}$$

and

$$\begin{aligned}&\int_{\mathcal{X}^n}|\delta^{(2)}(X^n)-\phi(p)|^2p(X^n)\nu(dX^n)\\ <&\int_{\mathcal{X}^n}|\delta^{(1)}(X^n)-\phi(p)|^2p(X^n)\nu(dX^n)\quad\text{for some } p\in\Gamma).\end{aligned}\tag{4.4.16}$$

In this case $\hat{p}_n^{(2)}$ ($\delta^{(2)}(X^n)$) is said to *dominate* $\hat{p}_n^{(1)}$ ($\delta^{(1)}(X^n)$). If no such uniformly dominating estimator exists, then $\hat{p}_n^{(1)}$ ($\delta^{(1)}(X^n)$) is said to be *admissible*. The following proposition is a consequence of Theorems 4.4.1 to 4.4.3.

Theorem 4.4.4 *The stochastic complexity density estimator $\tilde{p}_n$ ($\phi(\tilde{p}_n)$) is admissible for the estimation of a density p ($\phi(p)$) in the class Γ.*

4.5 Completeness

Completeness is another optimum property related to admissibility. Let D be the class of all density decision rules δ with finite risk $R(p,\delta)$ for any $p\in\Gamma=\{p_1,p_2,\cdots\}$ which consists of at most countable number of densities. A class of decision rules

$D' \subseteq D$ is said to be *complete* if given any rule $\delta \in D$ not in D' there exists a rule $\delta_0 \in D'$ that dominates δ. D' is said to be *minimal complete* if D' is complete and no subclass of D' is complete.

To achieve the main result in this section we assume there exists a complete class $D_1 \subseteq D$ and a positive constant $m > 1$ satisfying $\sum_{j=1}^{\infty} m^{-j} R(p_j, \delta) < \infty$ for any $\delta \in D_1$. This assumption implies that any subsequence of $\{R(p_j, \delta)\}$ could tend to infinite but at a restricted rate, by which it will be seen that we can define a metric in a related space to facilitate our mathematical proof.

Now we consider the set S_1, defined as

$$S_1 = \{\boldsymbol{y} = (y_1, y_2, \cdots) \mid \text{for some } \delta \in D_1, y_j = R(p_j, \delta) \text{ for } j = 1, 2, \cdots\}. \quad (4.5.1)$$

It is easily seen that $\sum_{j=1}^{\infty} m^{-j} y_j < \infty$ for any $\boldsymbol{y} \in S_1$. Denote by D_0 the class of all stochastic complexity density estimators in D_1 and let

$$S_0 = \{\boldsymbol{y} = (y_1, y_2, \cdots) \mid \text{for some } \delta \in D_0, y_j = R(p_j, \delta) \text{ for } j = 1, 2, \cdots\}. \quad (4.5.2)$$

Clearly S_0 is nonempty and $S_0 \subseteq S_1$. S_1 and S_0 can be transformed to obtain two subsets of l^1, where l^1 denotes the space of all sequences $\{x_n\}$ of points of $\mathcal{R}^{\infty}$ such that $\sum_{n=1}^{\infty} |x_n| < \infty$, i.e.

$$S_1(m) = \{\boldsymbol{y} = (y_1, y_2, \cdots) \mid \text{for some } \delta \in D_1, y_j = m^{-j} R(p_j, \delta) \text{ for } j = 1, 2, \cdots\} \quad (4.5.3)$$

and

$$S_0(m) = \{\boldsymbol{y} = (y_1, y_2, \cdots) \mid \text{for some } \delta \in D_0, y_j = m^{-j} R(p_j, \delta) \text{ for } j = 1, 2, \cdots\}. \quad (4.5.4)$$

Let S_1^* be the convex hull of S_1 defined as the set of all finite convex linear combination of the points of S_1, that is

$$S_1^* = \{\boldsymbol{z} : \boldsymbol{z} = \sum_{i=1}^{k} \lambda_i \boldsymbol{y}_i, \boldsymbol{y}_i \in S_1, \lambda_i > 0, \sum_{i=1}^{k} \lambda_i = 1\}. \quad (4.5.5)$$

Similarly we can get $S_1^*(m)$ as the convex hull of $S_1(m)$ and a subset of l^1. We know that if S is a convex set, then the closure of S is convex, and the intersection of two convex sets is also convex.

For the discussion of the completeness of the stochastic complexity estimators we need some concepts from Ferguson (1967, pp. 63-64).

A set S in the space l^1 is said to be *bounded from below* if there exists a finite number M such that for every $\boldsymbol{y} = (y_1, y_2, \cdots) \in S$, $y_j \geq -M$ for $j = 1, 2, \cdots$.

Let $\boldsymbol{x}$ be a point in l^1. The *lower quantant* at $\boldsymbol{x}$, denoted by $Q_{\boldsymbol{x}}$, is defined as the set

$$Q_{\boldsymbol{x}} = \{\boldsymbol{y} \in l^1; y_j \leq x_j \text{ for } j = 1, 2, \cdots\}. \tag{4.5.6}$$

A point $\boldsymbol{x}$ is said to be a *lower boundary point* of a convex set $S \subseteq l^1$ if $Q_{\boldsymbol{x}} \cap \bar{S} = \{\boldsymbol{x}\}$, where $\bar{S}$ is the closure of S. The set of lower boundary points of a convex set S is denoted by $\lambda(S)$.

A convex set $S \subseteq l^1$ is said to be *closed from below* if $\lambda(S) \subset S$.

Lemma 4.5.1 *If a nonempty convex set $S \subset l^1$ is bounded by 0 from below, then $\lambda(S)$ is not empty.*

Proof: Let $w_1, w_2, \cdots$ be a sequence of positive numbers satisfying $\sum_{i=1}^{\infty} w_i = 1$, and let T denote the set of all numbers of the form $t = \sum_{j=1}^{\infty} w_j y_j$, where $\boldsymbol{y} = (y_1, y_2, \cdots) \in S$.

$$T = \{t = \sum_{j=1}^{\infty} w_j y_j \text{ for some } \boldsymbol{y} \in S\}. \tag{4.5.7}$$

T is bounded by 0 from below because S is bounded by 0 from below. Let $t_0 = \inf\{t : t \in T\}$ and let $\boldsymbol{y}^{(n)} \in S$ be a sequence of points for which $\sum_{j=1}^{\infty} w_j y_j^{(n)} \to t_0$. Since $w_j > 0$ it follows that each sequence $y_j^{(n)}$ is bounded from above. Thus, using the principle of diagonal selection we can find a subsequence $\boldsymbol{y}^{(n_j)}$ of $\boldsymbol{y}^{(n)}$ with a finite limit which converges coordinatewise to a point $\boldsymbol{y}^0$ for which $\sum_{j=1}^{\infty} w_j y_j^0 = t_0$. Therefore $\boldsymbol{y}^0$ is a limit point under the l^1 metric $\| \ \|_1$.

Now we show that $\boldsymbol{y}^0 \in \lambda(S)$. First we note that $\{\boldsymbol{y}^0\} \subset Q_{\boldsymbol{y}^0} \cap \bar{S}$ because $\boldsymbol{y}^0$ is a limit point of S, i.e. $\boldsymbol{y}^0 \in \bar{S}$. On the other hand $Q_{\boldsymbol{y}^0} \cap \bar{S} \subset \{\boldsymbol{y}^0\}$, for if $\boldsymbol{y}'$ is any point of $Q_{\boldsymbol{y}^0}$ other than $\boldsymbol{y}^0$ itself, then $\sum_{j=1}^{\infty} w_j y_j' < t_0$. This contradicts the assumption that t_0 is a lower bound of T. Thus $Q_{\boldsymbol{y}^0} \cap \bar{S} = \{\boldsymbol{y}^0\}$, implying that $\boldsymbol{y}^0 \in \lambda(S)$. Hence $\lambda(S)$ is not empty. □

Lemma 4.5.2 *If a nonempty convex set $S \subset l^1$ is bounded by 0 from below, then for any $\boldsymbol{x} \in S$ but not in $\lambda(S)$, there exists a point $\boldsymbol{y}$ in $\lambda(S)$ so that $y_j \leq x_j$ for $j = 1, 2, \cdots$ and $y_{j'} < x_{j'}$ for some j'.*

Proof: Suppose $\boldsymbol{x} \in S$ but $\boldsymbol{x} \notin \lambda(S)$. Because S is convex $\bar{S}$ is also convex, thus $S' = Q_{\boldsymbol{x}} \cap \bar{S}$ is convex too and nonempty. S' is bounded by 0 from below, for S is bounded by 0 from below. By Lemma 4.5.1 $\lambda(S')$ is nonempty. Let $\boldsymbol{y} \in \lambda(S')$, then by definition $\{\boldsymbol{y}\} = Q_{\boldsymbol{y}} \cap \overline{S'}$. Furthermore, $\boldsymbol{y} \in Q_{\boldsymbol{x}}$ since $\boldsymbol{y} \in \overline{S'} = \overline{Q_{\boldsymbol{x}} \cap \bar{S}} \subset \bar{Q}_{\boldsymbol{x}} = Q_{\boldsymbol{x}}$. Finally, $\boldsymbol{y} \in \lambda(S)$ because $\{\boldsymbol{y}\} = Q_{\boldsymbol{y}} \cap \overline{S'} = Q_{\boldsymbol{y}} \cap \overline{Q_{\boldsymbol{x}} \cap \bar{S}} = Q_{\boldsymbol{y}} \cap Q_{\boldsymbol{x}} \cap \bar{S} = Q_{\boldsymbol{y}} \cap \bar{S}$. Now, we know that $\boldsymbol{y} \in Q_{\boldsymbol{x}} - \{\boldsymbol{x}\}$, hence $y_j \leq x_j$, $j = 1, 2, \cdots$ and at least for some j', $y_{j'} < x_{j'}$. □

By Theorem 2.5 of Valentine (1964, pp. 22) any hyperplane H in space l^1 can be expressed as $H = [f : \kappa]$, where f is a linear functional nonidentically zero on l^1, κ is a real constant and $[f : \kappa]$ denotes the set of all points $\boldsymbol{x} \in l^1$ for which $f(\boldsymbol{x}) = \kappa$. The hyperplane H bounds a set $S \subset l^1$ if either $f(S) \geq \kappa$ or $f(S) \leq \kappa$ holds, and H separates two sets U and V in l^1 if either $f(U) \geq \kappa$, $f(V) \leq \kappa$ or $f(U) \leq \kappa$,$f(V) \geq \kappa$ holds.

Lemma 4.5.3 *(Valentine, 1964, pp. 25) A hyperplane $H = [f : \kappa]$ in l^1 bounds a nonempty open set if and only if f is continuous with $f \not\equiv 0$.*

Lemma 4.5.4 *(DeVito, 1978, pp. 42-43) The vector space of all linear continuous functionals on l^1 is equivalent to l^∞, where l^∞ denotes the space of all sequences $\{x_n\}$ of $\mathcal{R}^\infty$ such that $\sup\{|x_n| \mid n = 1, 2, \cdots\}$ is finite.*

Lemma 4.5.5 *(Separation Theorem (Valentine, 1964, pp. 24) Suppose U and V are two nonempty convex subsets of a linear space $\mathcal{L}$. Also suppose the interior of U is nonempty and that $V \cap intU = \emptyset$. Then there exists a hyperplane H which separates U and V.*

Lemma 4.5.6 *(Ferguson, 1967, pp. 55) If a set of decision rules D' is a complete class, it must contain all admissible rules.*

Theorem 4.5.1 *If $S_1(m)$ and $S_1^*(m)$ are defined as above, S_1^* is bounded by 0 from below and closed from below, then $\lambda(S_1^*(m)) \subset S_0(m)$.*

Proof: Because $S_1^*(m)$ is also closed from below, $\lambda(S_1^*(m)) \subset S_1^*(m)$. If $\boldsymbol{x} \in \lambda(S_1^*(m))$, then $\{\boldsymbol{x}\} = Q_{\boldsymbol{x}} \cap \overline{S_1^*(m)} = Q_{\boldsymbol{x}} \cap S_1^*(m)$. Thus, $Q_{\boldsymbol{x}} - \{\boldsymbol{x}\}$ and $S_1^*(m)$ are disjoint convex sets. By the Separation Theorem, their exists a hvperplane H which separates $Q_{\boldsymbol{x}} - \{\boldsymbol{x}\}$ and $S_1^*(m)$ and by Lemma 4.5.3 $H = [f : \kappa]$, where κ is a real constant and f is a nonidentically zero continuous linear functional on l^1. From Lemma 4.5.4 it follows that there exists $\{\beta_i\}$ satisfying $\sup_i |\beta_i| < \infty$ such that

$$f(\boldsymbol{x}) = \sum_{i=1}^{\infty} \beta_i x_i = \boldsymbol{\beta}^T \boldsymbol{x} \tag{4.5.8}$$

for any $\boldsymbol{x} \in l^1$. Thus $\boldsymbol{\beta}^T \boldsymbol{y} \leq \boldsymbol{\beta}^T \boldsymbol{z}$ for any $\boldsymbol{y} \in Q_{\boldsymbol{x}} - \{\boldsymbol{x}\}$ and $\boldsymbol{z} \in S_1^*(m)$. If one of the coordinates β_j of $\boldsymbol{\beta}$ were negative, then by choosing $\boldsymbol{y}$ so that y_j is sufficiently negative, we would have $\boldsymbol{\beta}^T \boldsymbol{y} > \boldsymbol{\beta}^T \boldsymbol{x}$. Hence $\beta_j \geq 0$ for all j. By the continuity of f

$$\boldsymbol{\beta}^T \boldsymbol{x} \leq \boldsymbol{\beta}^T \boldsymbol{z} \quad \text{will hold for all } \boldsymbol{z} \in S_1^*(m). \tag{4.5.9}$$

By the definition of $S_1^*(m)$, there exist $\lambda_1, \lambda_2, \cdots, \lambda_k$ with $\lambda_i > 0$ and $\sum_{i=1}^k \lambda_i = 1$ so that $\boldsymbol{x} = \sum_{i=1}^k \lambda_i \boldsymbol{y}_i$ where $\boldsymbol{y}_i \in S_1(m)$. From (4.5.9)

$$\sum_{i=1}^{k} \lambda_i \boldsymbol{\beta}^T \boldsymbol{y}_i \leq \boldsymbol{\beta}^T \boldsymbol{z} \quad \text{for all } \boldsymbol{z} \in S_1(m), \tag{4.5.10}$$

which implies

$$\boldsymbol{\beta}^T \boldsymbol{y}_i = \min_{\boldsymbol{z} \in S_1(m)} \boldsymbol{\beta}^T \boldsymbol{z} \quad \text{for } i = 1, 2, \cdots, k. \tag{4.5.11}$$

From

$$\boldsymbol{\eta}^T \boldsymbol{y}_i^* = \min_{\boldsymbol{z} \in S_1} \boldsymbol{\eta}^T \boldsymbol{z} \quad \text{for } i = 1, 2, \cdots, k \tag{4.5.12}$$

where $\eta_j = m^{-j} \beta_j$ and $y_{ij}^* = m^j y_{ij}$, $j = 1, 2, \cdots$, $i = 1, 2, \cdots, k$, it follows that $\sum_{j=1}^{\infty} \eta_j \leq \sup_j |\beta_j| < \infty$ and $\boldsymbol{y}_i^* \in S_1$ for $i = 1, 2, \cdots, k$. Normalizing $\boldsymbol{\eta}$ by letting $\eta_j^* = \eta_j / \sum_{j=1}^{\infty} \eta_j$, we have

$$\boldsymbol{\eta}^{*T} \boldsymbol{y}_i^* = \min_{\boldsymbol{z} \in S_1} \boldsymbol{\eta}^{*T} \boldsymbol{z} \quad \text{for } i = 1, 2, \cdots, k. \tag{4.5.13}$$

So by Theorem 4.4.1 $\boldsymbol{y}_1^* = \boldsymbol{y}_2^* = \cdots = \boldsymbol{y}_k^* = \boldsymbol{y}^* \in S_0$ and hence $\boldsymbol{y}_1 = \boldsymbol{y}_2 = \cdots = \boldsymbol{y}_k = \boldsymbol{x} \in S_0(m)$. □

Theorem 4.5.2 *If, for a given decision problem* (Γ, D, R), *there exists a complete class* $D_1 \subseteq D$ *and a positive constant* $m > 1$ *satisfying* $\sum_{j=1}^{\infty} m^{-j} R(p_j, \delta) < \infty$ *for any* $\delta \in D_1$, *and* S_1^*, *the convex hull of* S_1 *defined above is bounded from below and closed from below, then* D_0 *is a minimal complete class and consists of exactly all the stochastic complexity density estimators.*

Proof: For any $\delta \in D_1$ not in D_0, let $y_j = m^{-j} R(p_j, \delta)$, $j = 1, 2, \cdots$ which implies that $\boldsymbol{y} \in S_1(m) \subset S_1^*(m)$ but not in $S_0(m)$. By Theorem 4.5.1 $y \notin \lambda(S_1^*(m))$. From Lemma 4.5.2 and Theorem 4.5.1 again, there exist a point $\boldsymbol{y}' \in \lambda(S_1^*(m)) \subset S_0(m)$ so that $y_j' \leq y_j$ for $j = 1, 2, \cdots$ and at least for some j', $y_{j'}' < y_{j'}$. This means that there exists a stochastic complexity density decision $\delta' \in D_0$ so that $y_j' = m^{-j} R(p_j, \delta')$ and δ' dominates δ. D_0 is therefore a complete class.

As a consequence of Theorem 4.4.4, every decision rule in D_0 is admissible. Hence no proper subclass of D_0 could be complete because (Lemma 4.5.6) every complete class must contain all admissible rules. This implies that D_0 consists of exactly the admissible rules and exactly all the stochastic complexity density estimators, and forms a minimal complete class. □

Remark: The condition of boundedness from below is not necessary since the definition of the risk function already implies it.

Chapter 5

Stochastic Complexity in Histograms and Testing Homogeneity

5.1 Introduction

In digital data-transmission systems, analogue input signals are first converted into digital form at the transmitter, then transmitted through a communication channel and finally reconstructed into analogue signals at the receiver. The resulting output is not identical with the input due to a quantization process in which the whole range of input amplitudes is divided into a finite number of amplitude sub-ranges at the transmitter, and the input amplitudes in each sub-range are converted into the same digits. This idea of quantization of the input signal can be transplanted and generalized naturally to the problem of estimating the probability distribution or density for an observed system.

Suppose we observe a finite data-string $X^n = X_1, X_2, \cdots, X_n$ from a system and we wish to describe the probability distribution of this data-string. Through a quantizer, the whole range of X^n is divided into a finite number of subintervals the widths of which can be either equal or unequal. In each subinterval we select a representative value, then each observation in X^n is replaced by a representative value which falls in the same subinterval as this observation. Thus the resulting quantized data-string

is encoded in a string of binary digits and transmitted through a communication channel, where it is decoded to provide the output. To make the code words uniquely decodable, the encoding system must be prefix.

From Chapter 1, the determination of the encoding system is equivalent to finding some kind of predictive probability density of the data generating system. In most cases, the underlying probability density is unknown and must be estimated. Fortunately, the quantizer in the data-transmission system gives us an access to an estimate of the unknown density. The density estimator can be used to construct an encoding system, and vice versa, under an encoding system the code words of the observed data-string X^n should be as small as possible so that the cost of transmitting the code words is small. This requirement is the key for a criterion to find the optimal quantizer.

It is possible to construct a histogram-type density estimation for X^n when the number of subintervals, their width and probabilities are given. For a fixed number of subintervals with fixed location, the probability of each subinterval can be determined by the maximum likelihood principle. The locations of the subintervals can also be determined by the maximum likelihood principle and a recursive method. After that, a temporary histogram density estimator is obtained from which a prefix code for X^n can be constructed. The optimal number of subintervals will generate the shortest code words and consequently an optimal histogram density estimate. The optimal description of X^n is then the code words under an optimal coding system.

The code words of X^n can be obtained by either a non-predictive or a predictive manner (see Chapter 3 of Rissanen (1989)). Even though the predictive coding requires more code words for the encoding of X^n, it enables the data-transmission system for self-adjustment and updating by using the latest observations.

In Section 5.2 below we first discuss an optimal quantization scheme of the data for optimal description which provides a system of recursive equations for determining the optimal locations of the subintervals in the histogram. Then both the idealized code length and the idealized predictive code length are given for the description of

X^n. Finally, uniform almost sure asymptotic expansion and the almost sure lower and upper bounds for both code lengths are derived and the results are list in Theorem 5.2.2 to Theorem 5.2.4.

In Hall and Hannan (1988) and Yu and Speed (1992), the same type of stochastic complexity based histogram estimation is considered under the assumption of equal subinterval widths. Our results agree with that of Yu and Speed (1992) when this assumption applies.

As an application of stochastic complexity for optimal data description, in Section 5.3 we consider the problem of testing of homogeneity, i.e. the testing of the hypothesis that several independent samples are generated from the same population. A test procedure is proposed in which we use the difference of shortest predictive code lengths under the null and the alternative hypotheses respectively as a universal test statistics. The size of the test procedure is shown to be determined by the part of the code lengths which is used to describe the parameters in the histogram densities. The asymptotic power of the test procedure is shown to be 1.

5.2 Data Compression for Optimal Information Description

Suppose $X^n = X_1, X_2, \cdots, X_n$ is a simple random sample from an unknown density function f on $[s,t]$, where s, t are finite real numbers. If f were known, the description of the sample could be accomplished by constructing a string of predictive binary codes for X^n under the information source determined by f (see Rissanen (1989)). In other words, the description of the sample is the same as finding a predictive probability density for the sample.

To estimate an unknown density f the most frequently used method is based on data compression: first quantize the data set by partitioning the interval $[s,t]$ into a sequence of subintervals and then construct a histogram on the partition. The choice of the partition and the estimate of the probability for each subinterval may be determined by the maximum likelihood method if a fixed number of subintervals

are assumed.

Let $q^m = q_{0,m}, q_{1,m}, \cdots, q_{m,m}$ denote an increasing sequence of numbers, partitioning the interval $[s,t]$ into m subintervals $[q_{0,m}, q_{1,m}]$, $(q_{1,m}, q_{2,m}]$, $\cdots$, $(q_{m-1,m}, q_{m,m}]$, written as $Q_{1,m}, Q_{2,m}, \cdots, Q_{m,m}$, where $q_{0,m} = s$, $q_{m,m} = t$ and m is a fixed integer satisfying $m \leq n$. Denote $r_{i,m} = q_{i,m} - q_{i-1,m}$ as the length of $Q_{i,m}$ and $r = t - s$, the range of X^n. Consider the histogram densities defined by

$$f(x|p^m, q^m, s, t) = \sum_{i=1}^{m} \frac{p_{i,m}}{r_{i,m}} I_{Q_{i,m}}(x) \tag{5.2.1}$$

where $p^m = p_{1,m}, p_{2,m}, \cdots, p_{m,m}$ denotes a sequence of nonnegative parameters with sum unit, and $I_{Q_{i,m}}$ is the usual indicator function. The set of densities of the form (5.2.1) is denoted by H_m.

With the above notations, the log-likelihood function of the sample X^n under H_m is

$$\begin{aligned} L(X^n; H_m) &= \sum_{j=1}^{n} \log \left(\sum_{i=1}^{m} \frac{p_{i,m}}{r_{i,m}} I_{Q_{i,m}}(X_j) \right) \\ &= \sum_{i=1}^{m} n_{i,m} \log \frac{p_{i,m}}{r_{i,m}} \end{aligned} \tag{5.2.2}$$

where $n_{i,m} = \sum_{j=1}^{n} I_{Q_{i,m}}(X_j)$ is the number of data points falling into $Q_{i,m}$. (All logarithms are in base 2 throughout this chapter unless stated otherwise.) Since $n_{i,m}$ may be zero, the corresponding $p_{i,m}$ can not be optimized through maximization of $L(X^n; H_m)$, and the log-likelihood function needs to be modified to overcome that difficulty. This may be done by introducing m numbers $y_1, y_2, \cdots, y_m$ (abbreviated as y^m), where y_i is regarded as an observation from the uniform distribution on $Q_{i,m}$, and blending them thoroughly with the n observations X^n as if both y^m and X^m were generated from the same distribution. Then the log-likelihood function of X^n and y^m is

$$L_1(X^n; H_m) = \sum_{i=1}^{m} (n_{i,m} + 1) \log \frac{p_{i,m}}{r_{i,m}} \tag{5.2.3}$$

which does not depend on the particular values of y^m, and can, therefore be regarded as the log-likelihood function of X^n.

Applying the maximum likelihood principle the optimal partition q^m and probabilities p^m for a fixed m are the ones which maximize $L_1(X^n, H_m)$ subject to the conditions that $\sum p_{i,m} = 1$ and $\sum r_{i,m} = r$. Denoting

$$F = \sum_{i=1}^{m}(n_{i,m}+1)\log\frac{p_{i,m}}{r_{i,m}} + \lambda_1(\sum_{i=1}^{m} p_{i,m} - 1) + \lambda_2(\sum_{i=1}^{m} r_{i,m} - r), \tag{5.2.4}$$

differentiating F with respect to p_i's and setting the derivatives equal to zero we have

$$\frac{\partial F}{\partial p_{i,m}} = \frac{n_{i,m}+1}{p_{i,m}}\log e + \lambda_1 = 0, \qquad i = 1, 2, \cdots, m \tag{5.2.5}$$

from which $p_{i,m} = (n_{i,m}+1)/(n+m)$. Differentiating F with respect to p_i's twice, the resulting second derivative matrix

$$\left(\frac{\partial^2 F}{\partial p_{i,m}\partial p_{j,m}}\right) = (\log e)diag\left(-\frac{n_{1,m}+1}{p_{1,m}^2}, \cdots, -\frac{n_{m,m}+1}{p_{m,m}^2}\right) \leq 0.$$

Therefore a necessary condition for the maximization of (5.2.4) is that the probabilities $p_{i,m}$ are equal to the relative frequency $(n_{i,m}+1)/(n+m)$.

Since the allocation of $n_{i,m}$'s depends on the partition q^m, so are the ranges $r_{i,m}$'s. The function F is not continuous with respect to $r_{i,m}$'s unless the allocation of $n_{i,m}$'s is fixed. Under such allocation the local extreme value of $L_1(X^n; H_m)$ is achieved or approached when the $r_{i,m}$'s tend to their boundary values, since all $Q_{i,m}$'s, except $Q_{1,m}$, are half-closed half-open intervals. In order to keep the code length needed to describe the model short, we impose the restriction that the end points of every subinterval $Q_{i,m}$, except the two end points s and t, i.e. the sequence of break points $q_{1,m}, \cdots, q_{m-1,m}$, should be at least d units away from the nearest observations, where $d > 0$ is half of the precision of X^n. In other words, if the locations of the sample X^n are expressed in an ascending order $z^N = z_1 < z_2 < \cdots < z_N$, where $N \leq n$ because of possible ties, then q^m is a subsequence of the $(2N+2)$ long sequence

$$s,\ z_1 - d,\ z_1 + d,\ z_2 - d,\ z_2 + d, \cdots, z_N - d,\ z_N + d,\ t,$$

denoted as $s(X^n) = s_1, s_2, \cdots, s_{2N+2}$, with $q_{0,m} = s$ and $q_{m,m} = t$, such that the selected q^m have the largest likelihood $L_1(X^n; H_m)$ among all the selections.

There are $\begin{pmatrix} 2N \\ m-1 \end{pmatrix}$ different selections for q^m within which the optimal sequence is to be found. In the following we provide a recursive method for finding the optimal q^m as well as the associated maximum likelihood values. A similar technique is used in Rissanen et. al. (1992). Let

$$\begin{aligned} L_1^*(X^n; m) &= \max_{q^m \subset s(X^n)} L_1(X^n; H_m) \\ &= \max_{q^m \subset s(X^n)} \sum_{i=1}^{m} (n_{i,m} + 1) \log \frac{n_{i,m} + 1}{(n+m) r_{i,m}}. \end{aligned} \tag{5.2.6}$$

It is easy to see that

$$\begin{aligned} L_1^*(X^{n(\tau)}; m) &= \max_{s_{m-1} \le q_{m-1,m} \in s(X^{n(\tau)})} \left\{ \max_{\{q_{1,m}, \cdots, q_{m-2,m}\} \in s(X^{n(q_{m-1,m})})} L_1(X^{n(q_{m-1,m})}; \right. \\ &\quad \left. H_{m-1}) + (n(\tau) - n(q_{m-1,m}) + 1) \log \frac{n(\tau) - n(q_{m-1,m}) + 1}{(n(\tau) + m) r_{m,m}} \right\} \\ &= \max_{s_{m-1} \le \nu \in s(X^{n(\tau)})} \left\{ L_1^*(X^{n(\nu)}; m-1) \right. \\ &\quad \left. + (n(\tau) - n(\nu) + 1) \log \frac{n(\tau) - n(\nu) + 1}{(n(\tau) + m) r_{m,m}} \right\} \end{aligned} \tag{5.2.7}$$

where $X^{n(\nu)}$ denotes the sequence of the observations falling within $[s, \nu]$, and $n(\nu)$ denotes the number of the observations in $X^{n(\nu)}$. The recursive equations (5.2.7) are to be solved for $m \ge 1$ and $\nu \in s(X^{n(\tau)})$ until the desired range includes all the observations. That is, the following maximum log-likelihood functions need to be solved in sequence

$$\begin{array}{llll} L_1^*(X^{n(s_2)}, 1), & L_1^*(X^{n(s_3)}, 1), & \cdots, & L_1^*(X^{n(\tau)}, 1), \\ L_1^*(X^{n(s_3)}, 2), & L_1^*(X^{n(s_4)}, 2), & \cdots, & L_1^*(X^{n(\tau)}, 2), \\ & \cdots & & \\ L_1^*(X^{n(s_{m+1})}, m), & L_1^*(X^{n(s_{m+2})}, m), & \cdots, & L_1^*(X^{n(\tau)}, m). \end{array} \tag{5.2.8}$$

for $m \le n$, where

$$L_1^*(X^{n(s_i)}, 1) = (n(s_i) + 1) \log \frac{1}{s_i - s}, \qquad 2 \le i \le 2N + 2$$

and

$$L_1^*(X^{n(s_k)}, k-1) = \sum_{i=1}^{k-1}(n(s_{i+1}) - n(s_i) + 1)\log\frac{n(s_{i+1}) - n(s_i) + 1}{(n(s_k) + k - 1)(s_{i+1} - s_i)},$$

for $2 \le k \le n+1$. For any fixed $m \le n$, the evaluation of (5.2.7) gives the maximum log-likelihood of X^n as well as the optimal partition $\{\tilde{Q}_{i,m}\}$ with about $m(4N + 3 - m)/2 \le 2m(n+2) - m^2/2$ operations. The corresponding optimal sequence of break points will be denoted by $\tilde{q}^m = \tilde{q}_{1,m}, \cdots, \tilde{q}_{m,m}$, and the widths of the subintervals by $\tilde{r}_{1,m}, \cdots, \tilde{r}_{m,m}$. In this chapter data quantization will always be based on the optimal partition $\{\tilde{Q}_{i,m}\}$ (except in the case of equal width quantization). For sake of simplicity the number of the data points falling into $\tilde{Q}_{i,m}$ will still be denoted as $n_{i,m} = \sum_{j=1}^{n} I_{\tilde{Q}_{i,m}}(X_j)$.

With an optimal procedure for the compression of the data, we are in a position to find the description of the data X^n.

Following Rissanen (1989), the description length of the data X^n, for fixed m and corresponding $\tilde{q}^m$, is defined as a two-part code length

$$-L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta) \tag{5.2.9}$$

where the first part $-L_1^*(X^n; m)$ can be interpreted as the code length needed to describe the data X^n under the given partition and histogram, and the second part L_2 is the code length needed to describe the functional form of the model employed. L_2 can be evaluated by first truncating the parameters m and $\tilde{q}^m$ to a prescribed precision δ and then encoding the resulting integers with the technique introduced in Elias (1975) and Rissanen (1989). Denote $\bar{a} = [a/\delta]$ as the nearest integer to a/δ, then

$$\begin{aligned} L_2(\tilde{q}^m, m, \delta) = \log &\begin{pmatrix} \sum_{i=1}^{m-1}\left|\overline{\tilde{r}_{i,m} - \frac{r}{m}}\right| + m - 2 \\ m - 2 \end{pmatrix} \\ &+ \log 2.865 + \log^*(\bar{m} + |\bar{s}| + \bar{r} + 1) + \\ &\log\frac{(\bar{m} + |\bar{s}| + \bar{r} + 3)!}{(\bar{m} + |\bar{s}| + \bar{r})!2!} + \log\frac{4!}{a_+!(3 - a_+)!} + |\log\delta|. \end{aligned} \tag{5.2.10}$$

Here $\log^*(a) = \log a + \log\log a + \cdots$, where the sum includes all the positive iterates, and a_+ is the number of nonnegative items in $\{\bar{m}, \bar{s}, \bar{r}\}$.

The length function (5.2.10) consists of three parts. Since the encoding of $\tilde{q}^m$ is equivalent to the encoding of $\tilde{r}_{1,m} - r/m, \cdots, \tilde{r}_{m-1,m} - r/m$, this will be achieved by a binary string beginning with $\overline{\tilde{r}_{1,m} - r/m}$ 0's and a 1, followed by $\overline{\tilde{r}_{2,m} - r/m}$ 0's and a 1, and so on until $\overline{\tilde{r}_{m-1,m} - r/m}$ 0's being added, but without attaching a 1 at the end, provided that m, s,t and d are given. Under this non-prefix encoding procedure the first term of (5.2.10) gives the code length of $\tilde{q}^m$. The second to the fifth terms of (5.2.10) are the code length needed for encoding $\bar{m}$, $\bar{s}$ and $\bar{t}$ (equivalent to $\bar{m}$, $\bar{s}$ and $\bar{r}$) in a prefix manner. In general we can encode a set of integers $\{\theta_1, \cdots, \theta_b\}$ in a prefix manner with about

$$L_3(\theta_1, \cdots, \theta_b) = \log 2.865 + \log^*(\theta + 1) + \log \frac{(\theta + b)!}{\theta!(b-1)!} + \log \frac{(b+1)!}{b_+!(b-b_+)!}$$

bits (Section 1.3). Here $\theta = \sum_i |\theta_i|$, and b_+ is the number of nonnegative items in $\{\theta_1, \cdots, \theta_b\}$. The last term gives us the code length for encoding the truncation precision δ. Since a_+ equals either 2 or 3, the fifth term of (5.2.10) can be replaced by 1 reflecting the fact that one digit is needed to tell if $\bar{s}$ is negative or nonnegative.

With the description length defined by (5.2.9) the shortest code length for the data X^n by the above encoding procedure is

$$\begin{aligned} &\min_m \{-L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta)\} \\ &\quad = -\sum_{i=1}^{m^*} (n_{i,m^*} + 1) \log \frac{n_{i,m^*} + 1}{(n + m^*)\tilde{r}_{i,m^*}} + L_2(\tilde{q}^{m^*}, m^*, \delta) \end{aligned} \tag{5.2.11}$$

where the minimization is done by searching for an optimal integer $m^* \leq n$ and δ is a prescribed precision.

If the sequence of break points are distributed uniformly in the interval $[s, t]$, then $\tilde{r}_{i,m} = r/m$ and the first term of (5.2.10) becomes zero. The expression (5.2.11) turns out to be

$$\min_m \left\{ -\sum_{i=1}^{m} (n_{i,m} + 1) \log \frac{(n_{i,m} + 1)m}{(n+m)r} + L_3(\bar{m}, \bar{s}, \bar{r}) + |\log \delta| \right\}. \tag{5.2.12}$$

An alternative to (5.2.11) is to use the idea of shortest predictive code length. This idea involves the ordering of the data X^n, either by location or by time of arrival, then finding the histogram density estimate based on the past and making appropriate modifications each time a new observation comes (Rissanen (1989) and Yu and Speed (1992)). In our situation the data X^n is ordered by location, as $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. For any fixed $m \leq n$, an optimal sequence of break points $\tilde{q}^m$ is obtained by solving the recursive equation (5.2.7). Let $i(X_{(j)})$ be the unique integer i such that $X_{(j)} \in \tilde{Q}_{i,m}$, and $n_{i,m}(\nu) = \sum_{X_l \leq \nu} I_{\tilde{Q}_{i,m}}(X_l)$ be the number of those X_l's satisfying $X_l \leq \nu$ and falling into the i-th subinterval $\tilde{Q}_{i,m}$. The histogram density estimate based on the first j observations $X_{(1)}, \cdots, X_{(j)}$ can be written as

$$\tilde{f}(x \mid X_{(1)}, \cdots, X_{(j)}, m) = \sum_{i=1}^{m} \frac{n_{i,m}(X_{(j)}) + 1}{(j+m)\tilde{r}_{i,m}} I_{\tilde{Q}_{i,m}}(x) \tag{5.2.13}$$

and the likelihood function of X^n can be constructed in a predictive manner as

$$\begin{aligned} \tilde{f}(X^n; m) &= \prod_{j=1}^{n} \tilde{f}(X_{(j)} \mid X_{(1)}, \cdots, X_{(j-1)}, m) \\ &= \prod_{j=1}^{n} \frac{n_{i(X_{(j)}),m}(X_{(j-1)}) + 1}{(j-1+m)\tilde{r}_{i(X_{(j)}),m}} \\ &= \frac{(m-1)!}{(n+m-1)!} \prod_{i=1}^{m} \frac{n_{i,m}!}{\tilde{r}_{i,m}^{n_{i,m}}}. \end{aligned} \tag{5.2.14}$$

In Rissanen et.al. (1992) $-\log \tilde{f}(X^n; m)$ is defined as the stochastic complexity of X^n under the given partition. Now the shortest predictive code length for the data X^n is

$$\begin{aligned} &\min_m \left\{ -\log \tilde{f}(X^n; m) + L_2(\tilde{q}^m, m, \delta) \right\} \\ &= -\sum_{i=1}^{\hat{m}} \log n_{i,\hat{m}}! + \sum_{i=1}^{\hat{m}} n_{i,\hat{m}} \log \tilde{r}_{i,\hat{m}} - \log \frac{(\hat{m}-1)!}{(n+\hat{m}-1)!} + L_2(\tilde{q}^{\hat{m}}, \hat{m}, \delta) \end{aligned} \tag{5.2.15}$$

where the minimization is achieved at $\hat{m} \leq n$ and δ is a prescribed precision. In particular, when the subintervals are of equal length, the expression (5.2.15) becomes

$$\min_m \left\{ n \log \frac{r}{m} + \log \binom{n}{n_{1,m}, \cdots, n_{m,m}} + \log \binom{n+m-1}{n} + L_3(\bar{m}, \bar{s}, \bar{r}) + |\log \delta| \right\}. \tag{5.2.16}$$

The relationship between the shortest code length (5.2.11) and shortest predictive code length (5.2.15) is established by the following results.

Theorem 5.2.1 *Let X^n be a simple random sample from an unknown density function f on $[s,t]$. Suppose the following conditions are satisfied:*

(i). $0 < c_1 \le f \le c_2 < \infty$, *where c_1, c_2 are two constants;*

(ii). The number of subintervals m in the quantization of X^n satisfies

$$n^{\gamma_1} \le m \le n^{\gamma_2},$$

where γ_1 and γ_2 are two constants satisfying $0 < \gamma_1 < \gamma_2 < 1$;

(iii). The width $\tilde{r}_{i,m}$ of each optimal subinterval $\tilde{Q}_{i,m}$ satisfies

$$b_1 m^{-\alpha_1} \le \tilde{r}_{i,m} \le b_2 m^{-\alpha_2}$$

uniformly for integers m in $[n^{\gamma_1}, n^{\gamma_2}]$, where b_1, b_2, α_1, α_2 are constants satisfying $1 \le \alpha_1 < \frac{1}{2} + \frac{1}{2\gamma_2}$, and $\max\left\{0, 2\alpha_1 - \frac{1}{\gamma_2}\right\} < \alpha_2 \le 1$.

Then uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$, the difference between the shortest code length and the shortest predictive code length of X^n

$$-\log \tilde{f}(X^n; m) + L_1^*(X^n; m) = \alpha' m \log m + \frac{1}{2} m \log n + O(m) \quad a.s. \qquad (5.2.17)$$

where $-\frac{1}{2}\alpha_1 \le \alpha' \le -\frac{3}{2} + \alpha_1$.

Note that if the support of the density f is finite, then $\alpha_2 \le 1 \le \alpha_1$ is implied by the condition (iii). Another useful implication of (ii) and (iii) is $\alpha_2 \le 1 \le \alpha_1 < \frac{1}{\gamma_2} < \frac{1}{\gamma_1}$.

Theorem 5.2.2 *In addition to the conditions (i), (ii) and (iii) in Theorem 5.2.1, suppose that*

(iv). f is absolutely continuous with derivative $\dot{f}$ a.e. such that $|\dot{f}(x)| \le c_3$.

Then uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$

$$(1) \qquad -Am^{\alpha_1} + (\alpha_2 - \alpha_1) m \log m + o(nm^{-2\alpha_2} + m \log m)$$
$$\leq -L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n)$$
$$\leq (\alpha_1 - 1) m \log m + C_f n m^{-2\alpha_2} + o(nm^{-2\alpha_2} + m \log n) \quad a.s. \text{ (5.2.18)}$$

if either $\alpha_1 \neq 1$ *or* $\alpha_2 \neq 1$, *and*

$$-L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) = O(nm^{-2} + m \log n) \quad a.s. \quad (5.2.19)$$

if $\alpha_1 = \alpha_2 = 1$.

$$(2) \quad -Am^{\alpha_1} + \frac{1}{2} m \log n + \left(\alpha_2 - \frac{3}{2}\alpha_1\right) m \log m + o(nm^{-2\alpha_2} + m \log m)$$
$$\leq -\log \tilde{f}(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n)$$
$$\leq \frac{1}{2} m \log n + \left(2\alpha_1 - \frac{5}{2}\right) m \log m + C_f n m^{-2\alpha_2} + o(nm^{-2\alpha_2} + m \log n) \quad a.s. \text{ (5.2.20)}$$

if either $\alpha_1 \neq 1$ *or* $\alpha_2 \neq 1$, *and*

$$-\log \tilde{f}(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) =$$
$$\frac{1}{2} m \log \frac{n}{m} + C_f' n m^{-2} + O(nm^{-2} + m \log n) \quad a.s. \quad (5.2.21)$$

if $\alpha_1 = \alpha_2 = 1$. *Here* $\log f^n(X^n) = \prod_{j=1}^n f(X_j)$, $C_f = \frac{b_2}{24} \int_s^t \frac{\dot{f}^2}{f}$, $A > 0$ *is a constant and* C_f' *is a constant between* $\frac{C_f b_1}{b_2}$ *and* C_f.

From Rissanen (1989) we know that $-\log f^n(X^n)$, the so-called Shannon Complexity, represents the optimal code length of X^n if the underlying density f is known. Thus the equations (5.2.18) and (5.2.20) represent, respectively, the redundant code length when using the coding processes corresponding to (5.2.11) and (5.2.15). Therefore uniform minimax bounds for the shortest code length (5.2.11) and the shortest predictive code length (5.2.15) respectively can be constructed as follows.

Theorem 5.2.3 *Under the conditions of Theorem 5.2.1 and Theorem 5.2.2 and having either* $\alpha_1 \neq 1$ *or* $\alpha_2 \neq 1$, *we have*

$$-M_1(n^{\alpha_1 \gamma_2} + n^{\gamma_2} \log n)$$

$$\leq \min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-L_1^*(X^n;m)+L_2(\tilde{q}^m,m,\delta)\}+\log f^n(X^n)$$
$$\leq M_2 n^{\frac{1}{1+2\alpha_2}}(\log n)^{\frac{2\alpha_2}{1+2\alpha_2}} \quad a.s. \tag{5.2.22}$$

and

$$-M_3(n^{\alpha_1\gamma_2}+n^{\gamma_2}\log n)$$
$$\leq \min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\left\{-\log\tilde{f}(X^n;m)+L_2(\tilde{q}^m,m,\delta)\right\}+\log f^n(X^n)$$
$$\leq M_4 n^{\frac{1}{1+2\alpha_2}}(\log n)^{\frac{2\alpha_2}{1+2\alpha_2}} \quad a.s. \tag{5.2.23}$$

where M_1, M_2, M_3, M_4 are positive constants depending on f.

Finally we give a result for the special case of $\alpha_1 = \alpha_2 = 1$.

Theorem 5.2.4 *Under the conditions of Theorem 5.2.1 and Theorem 5.2.2 and $\alpha_1 = \alpha_2 = 1$, the following statements hold.*

$$(a)\quad \min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-L_1^*(X^n;m)+L_2(\tilde{q}^m,m,\delta)\}+\log f^n(X^n)$$
$$= O(n^{\frac{1}{3}}(\log n)^{\frac{2}{3}}) \quad a.s., \tag{5.2.24}$$
$$(b)\quad \min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\left\{-\log\tilde{f}(X^n;m)+L_2(\tilde{q}^m,m,\delta)\right\}+\log f^n(X^n)$$
$$= M_5 n^{\frac{1}{3}}(\log n)^{\frac{2}{3}} \quad a.s., \tag{5.2.25}$$
$$(c)\quad m^* = O((n/\log n)^{\frac{1}{3}}) \quad a.s., \tag{5.2.26}$$
$$(d)\quad \hat{m} = M_6(n/\log n)^{\frac{1}{3}} \quad a.s.. \tag{5.2.27}$$

where M_5, M_6 are positive constants depending on f.

The proofs of Theorem 5.2.1 to Theorem 5.2.4 will be presented in Section 5.4.

The equations (b) and (d) agree with (ii) and (iv) of Theorem 2.4 of Yu and Speed (1992). Note that even though the predictive code length (5.2.20) is longer than the code length (5.2.18) with an infinite number of digits as $n \to \infty$, both of them have the minimax bound of the same order. In addition, the estimates (b) and (d) are better than (a) and (c) respectively. The predictive encoding process is therefore preferable and will be the subject of study in the remainder of this chapter.

From Theorem 5.2.3 and Theorem 5.2.4 it follows that for variable-width subintervals the minimax bound for the predictive code lengths (5.2.15) is no better than for the uniform width subintervals (see Theorem 2.4 of Yu and Speed (1992)) - unless $\alpha_1 = \alpha_2 = 1$, i.e. $\tilde{r}_{i,m} = O(m^{-1})$ - in which the same order of the bound is achieved. This is somewhat surprising and discouraging. It suggests that even though the finite sample behavior of a variable-width subinterval histogram is very likely to be better than that of an equal-width one, the use of the former histogram density is recommended only when the optimal widths $\tilde{r}_{i,m}$'s are of order $O(m^{-1})$.

5.3 Hypothesis Testing for Homogeneity

5.3.1 Background

One of the basic problems in statistical inquiry is the two-sample problem of testing the equality of two distributions, and more generally, the k-sample problem of testing the homogeneity of the distributions of several populations ($k > 2$). A typical example, commonly referred to as the one-way layout problem, is the comparison of several of treatments with a control, where the hypothesis of no treatment effect is tested against the alternative of at least one effect.

Under a parametric setting when the normality of the populations is assumed, the appropriate test is based on Student's t for the problem of equal means of two populations. However, when approximate normality is suspected but not fully trusted, one may replace the t-test by its permutation analogue, which can again be approximated by a t-test. For the case of homogeneity of means of more than 2 populations, the appropriate F test is used which is based on the assumption of normality and a common variance of the populations, the latter of which is tested by some more or less robust tests like the classic Bartlett's test. For the case where the assumption of a common variance can not be maintained, the so-called generalized Behrens-Fisher problem, other tests have been proposed. For a review see Lehmann (1986b).

To achieve robustness against the violation of some of the assumptions of the

parametric tests one may consider nonparametric alternatives. Usually a distribution-free statistic which is based on the ranks of the observations, and satisfying some invariance principles, is constructed to test homogeneity. The two most familiar ones are the two-sample Wilcoxon test and the Kruskal-Wallis test. The theory of these and related rank tests can be found in Hájek and Šidák (1967), Lehmann (1975), Randles and Wolfe (1979), and Hettmansperger (1984), and others.

All the tests cited above require that the different populations have the same distributional shape with the difference only in the location or the scale parameter, which sometimes can be explained by an additive or multiplicative treatment effect or both. But seldom are these claims statistically tested. Moreover, while these tests are sensitive to the location or scale difference, they may not detect differences of other types. The most commonly employed Smirnov test (see, for example, Conover et al., 1971) is consistent against all types of differences that may exist among the k populations.

By using the data compression method developed in Section 5.2, we will argue that the principle of stochastic complexity and minimal description length (MDL) have important roles to play in testing the homogeneity of the k populations against any type of difference among them.

Suppose we are given a set of data consisting of k independent random samples: $X_{11}, X_{12}, \cdots, X_{1n_1}$ with size of n_1; $X_{21}, X_{22}, \cdots, X_{2n_2}$ with size of n_2, $\cdots$, and $X_{k1}, X_{k2}, \cdots, X_{kn_k}$ with size of n_k, $k \geq 2$ and all the observations are independent. Let $F_1(x), F_2(x), \cdots, F_k(x)$ represent, respectively, their unknown population distribution functions and $f_1(x), f_2(x), \cdots, f_k(x)$ their corresponding density functions. We are now interested in testing if these k distributions are identical against the alternative that some kind of difference exists among them.

Our test procedure operates as follows. First, an idealized code length, the stochastic complexity, based upon the class of histogram density estimators with equal-width bins is computed for each independent random sample, which, when minimized, gives the optimal number of the bins with the associated density estimator and the proper

measurement of the information contained in each sample (Hall and Hannan (1988), Rissanen et.al (1992)). Second, the same kind of stochastic complexity is computed for the pooled sample, which, when minimized gives the estimator of the associated mixed density. Finally, a comparison is made between the stochastic complexity of the pooled sample and the sum of the stochastic complexities of all the samples; if the former one is smaller then the hypothesis of homogeneity of the k distributions are accepted, the hypothesis is rejected otherwise.

The novelty of our approach lies in using the principle of minimum description length and stochastic complexity instead of the classic methods which employ the empirical distribution. A major drawback of the commonly used classic tests is that they may be applied only to samples of equal sizes. This is because tables for the case of unequal sample sizes are unavailable, and must be obtained individually in each case. From a practical standpoint, however, the required calculations could even overtax the capacity of a computer. Our proposed method removes this difficulty because (a) it does not require the knowledge of the distribution of the test statistic, and (b) the procedure is justified for all continuous distributions and all sample sizes. Furthermore, with this new method one does not need to choose the level of significance of the test, for it becomes defined automatically.

5.3.2 The Test Procedure

Let $(X_{11}, \cdots, X_{1n_1}), (X_{21}, \cdots, X_{2n_2}), \cdots$, and $(X_{k1}, \cdots, X_{kn_k})$ (abbreviated as $X_1^{n_1}$, $X_2^{n_2}$, $\cdots$, $X_k^{n_k}$) be k independent random samples with sizes $n_1, n_2, \cdots, n_k$, $\sum_{i=1}^{k} n_i = n$, and unknown population density functions $f_1(x), f_2(x), \cdots, f_k(x)$ respectively. The problem is that of testing the hypothesis

$$\begin{aligned} H_0 &: \quad f_1 = f_2 = \cdots = f_k \quad \text{against} \\ H_a &: \text{ at least two of them are not equal.} \end{aligned} \tag{5.3.1}$$

We begin the analysis by first establishing the information contained in each of these k samples. If the densities $f_1, f_2, \cdots, f_k$ are known, the Shannon's entropy (if

exists)

$$-\sum_{j=1}^{n_i}\int f_i(X_{ij})\log f_i(X_{ij})dX_{ij} = -n_i\int f_i\log f_i,$$

$i = 1,\cdots,k$, respectively, will give us the optimal mean code length for each sample.(In this chapter all logarithms are in base 2.) In this sense,

$$-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\int f_i(X_{ij})\log f_i(X_{ij})dX_{ij} = -\sum_{i=1}^{k}n_i\int f_i\log f_i$$

gives us a measurement of information contained in the k samples.

Suppose now that we mistakenly ignore the differences that may exist among the k density functions and encode the k samples of the data as if they were from a single information source. The mean code length then is

$$-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\int f_i(X_{ij})\log f_{mix}(X_{ij})dX_{ij} = -n\int f_{mix}\log f_{mix},$$

where $f_{mix} = \sum_{i=1}^{k}(n_i/n)f_i$ is a mixture density of $f_1,\cdots,f_k$.

The inequality $-\sum_{i=1}^{k}n_i\int f_i\log f_i \le -n\int f_{mix}\log f_{mix}$, which holds due to the convexity of $x\log x$, i.e.

$$-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\int f_i(X_{ij})\log f_i(X_{ij})dX_{ij} \le -\sum_{i=1}^{k}\sum_{j=1}^{n_i}\int f_i(X_{ij})\log f_{mix}(X_{ij})dX_{ij} \quad (5.3.2)$$

where equality holds if and only if all the densities $f_1,\cdots,f_k$ are equal (except a set of measure zero), suggests that if the data are encoded in two distinct ways, each sample separately as well as a pooled sample, and the resulting code length for the latter is found larger than that of the former, then the conclusion that the null hypothesis H_0 is violated may be warranted. Indeed, this makes sense because, following the arguments by Shannon (1948) and Rissanen (1989), the optimal mean code length per symbol is a bound which can only rarely be beaten by any other per symbol code length, refer to Theorem 1.2.1 and 1.2.2.

The principle of minimum description length (*MDL*) and the notion of stochastic complexity (Rissanen, 1989) point out the way to estimate the optimal length encoding of the data. Suppose the unknown densities f_i's belong to a parametric or a

nonparametric model class $\mathcal{M}$. To achieve the optimal encoding of a given sample, say $X_i^{n_i}$, we need to select a density $\hat{f}_i$ in $\mathcal{M}$ based on which the resulting length of the code for $X_i^{n_i}$, $-\log\left(\prod_{j=1}^{n_i}\hat{f}_i(X_{ij})\right)$, is as short as possible while at the same time $L(f)$, the code length for encoding $\hat{f}_i$ itself, is not too long. In other words, we select a density $\hat{f}_i$ for $X_i^{n_i}$ so that the resulting two-part code length achieves the following

$$\min_{f_i\in\mathcal{M}}\{-\log f_i(X_i^{n_i})+L(f_i)\},\quad i=1,2,\cdots,k \tag{5.3.3}$$

Similarly, if we combine the k samples together and encode the pooled sample, the resulting optimal code length will be

$$\min_{f_1,\cdots,f_k\in\mathcal{M}}\left\{-\log\prod_{i=1}^{k}f_{mix}(X_i^{n_i})+L(f_{mix})\right\} \tag{5.3.4}$$

There are some difficulties in performing the minimizations (5.3.3) and (5.3.4), because these are not directly computable from the data. To overcome these, we apply the so-called stochastic complexity based nonparametric histogram density estimator and compute the associated minimum description length of the data.

Suppose the data of each sample $X_i^{n_i}$, fall in the interval $[s_i,t_i]$, and the data of the combined k samples fall in the interval $[s,t]$, where $s=\min\{s_i,1\le i\le k\}$ and $t=\max\{t_i,1\le i\le k\}$. Let $\mathcal{M}_1$ be the class of histogram densities with equal-width bins, on which we shall demonstrate the minimizations (5.3.3) and (5.3.4). If we partition $[s_i,t_i]$ into m_i congruent subintervals C_{ij} for each sample, for $1\le j\le m_i$ and $1\le i\le k$, our histogram density estimator $\hat{f}_i(x)$ will take the value $(m_i/r_i)p_{ij}$ when $x\in C_{ij}$, where $r_i=t_i-s_i$ is the range, $p_{ij}\ge 0$ and $\sum_{j=1}^{m_i}p_{ij}=1$, $i=1,\cdots,k$. As in Hall and Hannan (1988) and Risannen, Speed and Yu (1992), we assume the uniform prior $\pi(\boldsymbol{p}_i)=(m_i-1)!$ on the simplex defined by $\boldsymbol{p}_i=(p_{i1},\cdots,p_{im_i})$ and evaluate the marginal likelihood of the sample $X_i^{n_i}$

$$\begin{aligned} l_i(X_i^{n_i};s_i,r_i,m_i) &= \int\prod_{j=1}^{n_i}\hat{f}_i(X_{ij})\pi(\boldsymbol{p}_i)d\boldsymbol{p}_i \\ &= \int\left(\frac{m_i}{r_i}\right)^{n_i}\left(\prod_{j=1}^{m_i}p_{ij}^{n_{ij}}\right)(m_i-1)!d\boldsymbol{p}_i \\ &= \left(\frac{m_i}{r_i}\right)^{n_i}\frac{(m_i-1)!\prod_{j=1}^{m_i}n_{ij}!}{(n_i+m_i-1)!} \end{aligned} \tag{5.3.5}$$

where n_{ij} denotes the number of the data points in sample $X_i^{n_i}$ that fall in the subinterval C_{ij}. Then the stochastic complexity, i.e. the abstract shortest code length for $X_i^{n_i}$ relative to the set of all histograms with fixed s_i, r_i and m_i, is given by

$$I(X_i^{n_i} \mid s_i, r_i, m_i) = -\log l_i(X_i^{n_i}; s_i, r_i, m_i)$$
$$= n_i \log \frac{r_i}{m_i} + \log \binom{n_i}{n_{i1}, \cdots, n_{im_i}} + \log \binom{n_i + m_i - 1}{n_i}, \qquad (5.3.6)$$
$$i = 1, \cdots, k.$$

where $\binom{n_i}{n_{i1}, \cdots, n_{im_i}} = \frac{n_i!}{\prod_j n_{ij}!}$ and $\binom{n_i + m_i - 1}{n_i} = \frac{(n_i+m_i-1)!}{n_i!(m_i-1)!}$. By the same argument, the stochastic complexity for the pooled sample $X^n = (X_1^{n_1}, X_2^{n_2}, \cdots, X_k^{n_k})$ relative to the set of all histograms, given $s, r(= b - s)$ and m equal-width bins, is

$$I(X^n \mid s, r, m) = n \log \frac{r}{m} + \log \binom{n}{n_{\cdot 1}, \cdots, n_{\cdot m}} + \log \binom{n + m - 1}{n}. \qquad (5.3.7)$$

Here, $n_{\cdot i}$ denotes the number of the data points in X^n that falls in the ith bin of the partitioned interval $[s, b]$.

Note that if $m_i > n_i$ (or $m > n$), there will always be some subintervals containing no observation. To describe the employed model we have to take some code length for the encoding of these unnecessary subintervals. This is hardly reasonable. Therefore we restrict in the class of histograms that

$$1 \le m_i \le n_i \quad (i = 1, \cdots, k) \text{ and } \quad 1 \le m \le n.$$

For the minimizations (5.3.3) and (5.3.4) we still need the code lengths required to encode the parameter sets $\{s_i, r_i, m_i, i = 1, \cdots, k\}$ and $\{s, r, m\}$, which will be combined, respectively, with (5.3.6) and (5.3.7) to provide us the data-based two-part code length corresponding to (5.3.3) and (5.3.4). Since the optimal m_i and m usually depend on the sample size, the code lengths needed to encode the parameters could be quite comparable to the stochastic complexities (5.3.6) and (5.3.7)—especially for small and medium sized samples—which would reduce the importance the stochastic complexity is playing in dominating the random structure of

the data. However, we can avoid such an unpleasant situation by truncating the number of decimal digits kept in the parameters, and encode instead the resulting $\{[s_i/10^d], [r_i/10^d], [m_i/10^d], i = 1, \cdots, k\}$ and $\{[s/10^d], [r/10^d], [m/10^d]\}$ as well as the optimized precision d, where $[y]$ denotes the nearest integer to y. (In the sequel we shall use $\bar{y}$ to denote $[y/10^d]$.) This means that the difference between each parameter and values within its neighborhood of width 10^d is ignored.

There is a natural restriction for the precision d that it ranges from minus the largest number of effective digits after decimal point of the observations to one less than the largest number of digits before the decimal point of the observations. For example, if the measurements of a given sample are all rounded to 3 decimals and the largest absolute value of the sample is 347.635, then d will be restricted within the interval [-3,2]. In the following analysis we assumed that d is given in advance.

Section 1.3 showed that for a set of integers $\{\theta_1, \theta_2, \cdots, \theta_b\}$ a prefix code can be found with about

$$L_3(\theta_1, \theta_2, \cdots, \theta_b) = \log 2.865 + \log^*(\theta) + \log \frac{(\theta + b)!}{\theta!(b-1)!} + \log \frac{(b+1)!}{b_+!(b-b_+)!} \qquad (5.3.8)$$

number of bits. Here $\theta = \sum_i |\theta_i|$, b_+ is the number of non-negative items in $\{\theta_1, \cdots, \theta_b\}$ and $\log^*(n) = \log n + \log\log n + \cdots$, where the sum includes all the positive iterates.

For a prescribed precision d we are now in a position to obtain the data-based expression of the idealized code length for a sample, which equals to

$$\min_{m_i}\{I(X_i^{n_i} \mid s_i, r_i, m_i) + L_3(\bar{s}_i, \bar{r}_i, \bar{m}_i) + |\log 10^d|\} \qquad (5.3.9)$$

for the sample $X_i^{n_i}$, $i = 1, \cdots, k$, and

$$\min_{m}\{I(X^n \mid s, r, m) + L_3(\bar{s}, \bar{r}, \bar{m}) + |\log 10^d|\} \qquad (5.3.10)$$

for the pooled sample X^n. Note that (5.3.9) and (5.3.10) exactly agree with (5.2.16).

Therefore, in the case of unequal densities $f_1, f_2, \cdots, f_k$, the idealized code length for encoding $X_1^{n_1}, \cdots, X_k^{n_k}$ is

$$\min_{m_1, \cdots, m_k} \left\{\sum_{i=1}^{k} I(X_i^{n_i} \mid s_i, r_i, m_i)\right.$$

$$
\begin{aligned}
&+L_3(\bar{s}_1,\bar{r}_1,\bar{m}_1,\cdots,\bar{s}_k,\bar{r}_k,\bar{m}_k)+|\log 10^d|\Big\} \\
= \min_{m_1,\cdots,m_k} &\left\{\sum_{i=1}^{k} n_i \log\frac{r_i}{m_i}+\sum_{i=1}^{k}\log\binom{n_i}{n_{i1},\cdots,n_{im_i}}\right. \\
&+\sum_{i=1}^{k}\log\binom{n_i+m_i-1}{n_i}+\log 2.865+\log^*\sum_{i=1}^{k}(|\bar{s}_i|+\bar{r}_i+\bar{m}_i) \\
&+\log\frac{(\sum_{i=1}^{k}(|\bar{s}_i|+\bar{r}_i+\bar{m}_i)+3k)!}{(\sum_{i=1}^{k}(|\bar{s}_i|+\bar{r}_i+\bar{m}_i))!(3k-1)!} \\
&\left.+\log\frac{(3k+1)!}{(3k)_+!(3k-(3k)_+)!}+|\log 10^d|\right\} \qquad (5.3.11)
\end{aligned}
$$

where $(3k)_+$ indicates the number of non-negative values in $\{s_1, r_1, m_1, s_2, r_2, m_2, \cdots, s_k, r_k, m_k\}$. Note that in (5.3.11) we use the shorter $L_3(\bar{s}_1,\bar{r}_1,\bar{m}_1,\cdots,\bar{s}_k,\bar{r}_k,\bar{m}_k)$ instead of the longer $\sum_{i=1}^{k} L_3(\bar{s}_i,\bar{r}_i,\bar{m}_i)$. The efficiency lies in the fact that the former length is obtained in a prefix manner.

By the theory of stochastic complexity, under the right probabilistic model (here the density function), or the right constraints inside the probabilistic pattern of the observations, the corresponding encoding process is expected to produce a shorter code length than the one corresponding to the wrong model, or one that ignores the right constraints in the underlying model. Therefore we could conclude that if the alternative hypothesis H_a is true the code length in (5.3.11) should be less than that in (5.3.10), since the encoding procedure corresponding to (5.3.10) is based on the wrong model stated in H_0. If, however, the null hypothesis is true, then (5.3.2) implies that both encoding procedures corresponding to (5.3.10) and (5.3.11) should give virtually the same code length. (5.3.10) on the other hand would more likely to result in smaller code length because in (5.3.11) one needs to encode more parameters. Clearly then, (5.3.10) and (5.3.11) can be used as test criteria to test H_0 against H_a in which the code lengths of the parameters play the role of determining the size of the test. Moreover, it enables us to go further and detect which of the k densities are different and which of them are identical by trying to beat the code length (5.3.11) by a more precise modeling of the data.

By Section 5.2, a more general encoding process, based on the histogram density with variable-width subintervals, can be applied to obtain the idealized code length for each data sample. Suppose the alternative hypothesis H_a; from (5.2.15) the total predictive code length needed for the k samples $X_1^{n_1}, X_2^{n_2}, \cdots, X_k^{n_k}$ is

$$\min_{m_1,\cdots,m_k} \left\{ -\sum_{i=1}^{k} \log \tilde{f}_i + \sum_{i=1}^{k} L_2(q_i^{m_i}, m_i, 10^d) \right\} \tag{5.3.12}$$

if we suppose the parameter truncation is based on the same precision d. Here $\tilde{f}_i = \tilde{f}_i(X_i^{n_i}; m_i)$ is the likelihood function of the i-th sample $X_i^{n_i}$ defined as (5.2.14), i.e.

$$\tilde{f}_i = \frac{(m_i - 1)!}{(n_i + m_i - 1)!} \prod_{j=1}^{m_i} \frac{n_{i,j,m_i}!}{\tilde{r}_{i,j,m_i}^{n_{i,j,m_i}}} \tag{5.3.13}$$

where $\tilde{r}_{i,j,m_i}$'s are the widths of the optimal partition $\{\tilde{Q}_{i,j,m_i}\}$ of the i-th sample. These are obtained by the maximum likelihood principle (5.2.4) for fixed number of subintervals m_i applied to the i-th sample. n_{i,j,m_i} is then the number of data points falling into the j-th subinterval $\tilde{Q}_{i,j,m_i}$.

Because all of the k samples are encoded simultaneously, the second term of (5.3.12) could be further reduced by a more efficient encoding process for the parameters $m_1, \cdots, m_k$, s and t defined as

$$L_4(\tilde{q}_1^{m_1}, \cdots, \tilde{q}_k^{m_k}, m_1, \cdots, m_k, \delta) = \sum_{i=1}^{k} \log \begin{pmatrix} \sum_{j=1}^{m_i - 1} \left| \tilde{r}_{i,j,m_i} - \frac{r}{m_i} \right| + m_i - 2 \\ m_i - 2 \end{pmatrix} + L_3(\bar{m}_1, \cdots, \bar{m}_k, \bar{s}, \bar{r}) + |\log 10^d| \tag{5.3.14}$$

where $\tilde{q}_i^{m_i}$ is the sequence of break points corresponding to the optimal partition $\{\tilde{Q}_{i,j,m_i}\}$. Therefore under the hypothesis H_a the total predictive code length (5.3.12) for the k samples can be replaced by a shorter code length

$$C(X_1^{n_1}, \cdots, X_k^{n_k}) = \min_{m_1,\cdots,m_k} \left\{ -\sum_{i=1}^{k} \log \tilde{f}_i(X_i^{n_i}; m_i) + L_4(\tilde{q}_1^{m_1}, \cdots, \tilde{q}_k^{m_k}, m_1, \cdots, m_k, 10^d) \right\} \tag{5.3.15}$$

where the minimum is attained at $\hat{m}_1, \cdots, \hat{m}_k$.

If the null hypothesis H_0 is true, that is the k samples $X_1^{n_1}, \cdots, X_k^{n_k}$ are drawn from the same population distribution, we can describe the information in the k samples using only the optimal code words required to encode the pooled sample $X^n = (X_1^{n_1}, \cdots, X_k^{n_k})$. By regarding the pooled sample X^n as drawn from a mixed distribution with density $f_{mix} = \sum_{i=1}^{k} \frac{n_i}{n} f_i$, the shortest predictive code length for encoding X^n is the one defined by (5.2.15):

$$\begin{aligned} C(X^n) &= \min_m \left\{ -\log \tilde{f}_{mix}(X^n; r) + L_2(\tilde{q}^m_{mix}, m, 10^d) \right\} \\ &= -\sum_{j=1}^{\hat{m}} \log n_{j,\hat{m}}! + \sum_{j=1}^{\hat{m}} n_{j,\hat{m}} \log r_{j,\hat{m}} - \log \frac{(\hat{m}-1)!}{(n+\hat{m}-1)!} + L_2(\tilde{q}^{\hat{m}}_{mix}, \hat{m}, 10^d) \quad (5.3.16) \end{aligned}$$

where the minimum is attained at $\hat{m}$.

The large sample asymptotic behavior of the discussed test procedure is presented in the following theorem.

Theorem 5.3.1 *Let $X_1^{n_1}, \cdots, X_k^{n_k}$ be simple random samples, respectively drawn from the unknown density functions $f_1, \cdots, f_k$ on $[s,t]$, and $X^n = (X_1^{n_1}, \cdots, X_k^{n_k})$ the pooled sample. Suppose that the conditions (i) to (iv) listed in Theorem 5.2.1 and Theorem 5.2.2 are satisfied for each $X_i^{n_i}$ and the corresponding f_i. Then the following statements hold.*

(i). If at least two of $f_1, \cdots, f_k$ are not equal almost everywhere, there exists a constant $\eta < 0$ such that

$$\frac{1}{n}[C(X_1^{n_1}, \cdots, X_k^{n_k}) - C(X^n)] < \eta \quad a.s. \qquad (5.3.17)$$

as $n_1 \to \infty, \cdots, n_k \to \infty$ satisfying $\frac{n_1}{n} > \varepsilon_1 > 0, \cdots, \frac{n_k}{n} > \varepsilon_k > 0$ for any set of prescribed constants $\varepsilon_1 \cdots, \varepsilon_k$.

(ii). If $f_1 = f_2 = \cdots = f_k$ a.s., then

$$\frac{1}{n}[C(X_1^{n_1}, \cdots, X_k^{n_k}) - C(X^n)] \to 0 \quad a.s. \qquad (5.3.18)$$

as $n_1 \to \infty, \cdots, n_k \to \infty$.

The proof of Theorem 5.3.1 will be given in section 5.4.

Since (5.3.10) and (5.3.11) are the special situation of (5.3.16) and (5.3.15) respectively, the above theorem is also true for the test procedure based on the encoding process for histograms with equal-width subintervals.

From part (i) of the above theorem we know that the asymptotic power of our test procedure is 1 in the limit as the sample sizes tend to infinity, i.e., almost surely, the shortest predictive code length under H_a is less than that under H_0 when H_a is true. However, when the null hypothesis H_0 is true, the difference of the two shortest predictive code lengths per observation tends to be zero almost surely as the sample sizes tend to be infinity. This implies that we need some threshold value for (5.3.18) to control the type I error when H_0 is true.

When the sample sizes are finite, the size of the test is essentially determined by the part of the code lengths required for encoding the parameters. In the encoding process corresponding to (5.3.15) there are more parameters ($\tilde{q}_1^{m_1}$, $\cdots$, $\tilde{q}_k^{m_k}$, m_1, $\cdots$, m_k, d) to be encoded than in the encoding process corresponding to (5.3.16) in which only $\tilde{q}_{mix}^{m}$, m and d are to be encoded. Thus the code length (5.3.15) is more likely to be larger than (5.3.16) if the null hypothesis H_0 is true. Furthermore, the size of the test varies with d, the precision truncating the parameters. If the parameters are not truncated, the code length used to describe the parameters will be quite comparable to that used to describe the data under the given histogram estimate, and the type I error will be quite small, while the type II error is likely to be large. On the other hand if the parameters are truncated too heavily, i.e. too much information suggested by the parameters is ignored, the type I error is likely to be large even though the type II error will be well controlled. As a rule of thumb, we use the precision of X^n as 10^d used to truncate the parameters. Exact formula for determining the optimal precision d is not available, but some heuristic grasp of how the power of the test varies with the precision d can be obtained by the simulation study later in this section. In Rissanen (1994a), Fisher information is used to find the stochastic complexity of a set of data coming from a parametric model class. As a consequence, there is no need at all to

choose the optimal precision to truncate the parameters. Some works on applying this idea to test the hypothesis of homogeneity is currently pursued by the author.

5.3.3 Two Examples

The first example uses the "PRO Football Scores" data of R. Lock (1992). In order to get an idea of how the criterion works, we compare only the pointspread (abbreviated as pts., Oddsmaker's points to handicap the favored team) data in the third week, the eighth week and the fourteenth week to assess the presence of a time shift in the scores.

Scores of the third week:

7.5	3.5	7.0	10.0	2.5	6.5	8.5	2.5	4.0	7.5	1.5	3.5
4.5	4.0	9.5	2.0	5.5	9.0	3.0	9.0	3.5	5.5	9.0	7.0
10.5	2.0	14.0	2.0	14.0	3.5	9.0	2.0	3.0	3.0	1.5	3.5
2.0	7.5	6.0	8.0	3.0	4.0						

Scores of the eighth week:

7.0	6.5	2.5	2.0	2.5	4.0	6.0	3.0	4.0	6.0	8.5
6.5	2.0	2.0	6.5	5.5	2.5	2.5	9.0	3.5	6.0	13.0
4.0	3.5	0.0	0.0	5.5	7.0	12.0	12.5	5.5	1.0	4.0
4.0	2.0	7.0	4.0	13.0						

Scores of the fourteenth week:

12.0	1.0	12.0	6.5	3.0	6.0	3.0	9.0	1.5	9.5	10.0
8.0	5.0	0.0	13.5	4.5	5.5	3.5	13.0	7.5	5.0	2.0
6.5	4.0	4.0	3.0	3.0	6.5	8.0	5.5	9.0	9.5	11.0
1.5	5.0	7.0	8.5	5.0	6.5					

with sample sizes $n_1 = 42, n_2 = 38, n_3 = 39$ respectively.

Under the null hypothesis of no time shift in the pointspread, the idealized code length (5.3.10) for the pooled sample with $m \leq 119$, is 363.27, and the corresponding optimal $m = 119$ and $d = 1$ (one less than the largest number of digits before the

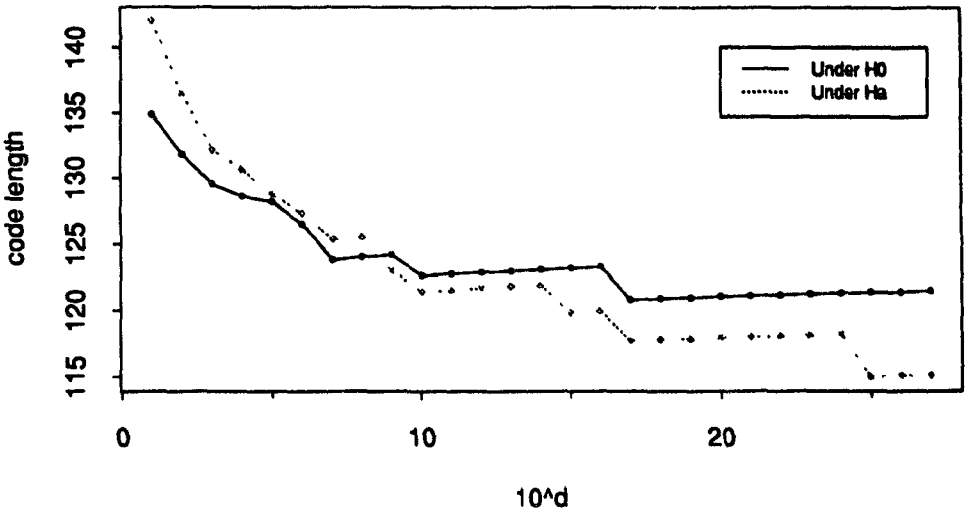
decimal point in the observations). Under the alternative hypothesis of some time effects in the pointspread, the idealized code length (5.3.11) for the three independent samples is 445.27, achieved at $m_1 = 5$, $m_2 = 2$, $m_3 = 1$ and $d = 1$. Because the idealized code length for the pooled sample is considerably smaller than that of the three independent samples, we conclude that there is no evidence of time effect in the pointspread, which concurs with the conclusion of the classic Student's t test for the mean difference of every two of the three samples. Figure 1 and Figure 2 show how the idealized code lengths of each individual sample and the pooled sample change with the number of bins employed in the corresponding histogram densities and with the precision d used to truncate the parameters.

In the second example, we generated two independent samples with sizes $n_1 = 15$ and $n_2 = 12$, respectively, from Gamma(4,3) and Uniform(4,18) distributions. The two samples are as follows

$$
\begin{aligned}
\boldsymbol{X}_1 = \; & 7.362 \;\; 8.876 \;\; 5.219 \;\; 10.506 \;\; 12.590 \;\; 9.552 \;\; 10.203 \;\; 11.144 \\
& 27.296 \;\; 3.105 \;\; 8.995 \;\; 4.955 \;\; 4.065 \;\; 10.822 \;\; 11.097 \\
\boldsymbol{X}_2 = \; & 6.645 \;\; 6.246 \;\; 7.589 \;\; 4.563 \;\; 11.131 \;\; 4.371 \;\; 6.743 \;\; 16.647 \\
& 15.412 \;\; 6.202 \;\; 15.134 \;\; 6.951
\end{aligned}
$$

Under the null hypothesis H_0 that there is no difference between the distributions which generated the two samples, the idealized code length (5.3.10) is 120.78 with $m = 3$ and $d = \log_{10} 17$, while under the alternative hypothesis H_a that there is a difference between the two distributions, the idealized code length (5.3.11) is 115.00 with $m_1 = 3, m_2 = 7$ and $d = \log_{10} 25$. The difference is clearly indicated by (5.3.10) and (5.3.11), but neither the classic Student's t test, which gives the p-value=0.7026, nor the Smirnov test, which is not significant at $\alpha = 0.2$, would indicate that difference. It is also interesting to note that (5.3.10) is always minimized at $m = 3$ when d is chosen to be from 0, $\log_{10} 2$, $\log_{10} 3$, $\cdots$, $\log_{10} 27$, while (5.3.11) is always minimized at $m_1 = 3$ and $m_2 = 7$. Figure 3 illustrates the relationship among (5.3.10), (5.3.11) and d.

Figure3: Relationship between Code Length and Precision



5.3.4 Simulation Studies

In this subsection we assess the finite sample performance of the proposed test procedure, based on the encoding process for histograms with equal-width subintervals, by a simulation study. We compare our method with the two sample t-test and the Smirnov test for equal and unequal sample sizes. The comparisons are in terms of the power of the test and based on 1000 repetitions. The results are summarized in Table 5.1 and Table 5.2.

Instead of using the optimal precision we choose some different but reasonable precision to truncate the parameters. It is found that there usually exists a precision d which makes both type I and II errors reasonably small.

The tables illustrate the following findings:

(i). When the samples are generated from normal distributions, the three tests are all efficient if the difference of the populations is the result of a mean shift. Both the Smirnov test and the stochastic complexity test are efficient when the difference is the result of a change in the variance, but the latter is better.

(ii). When the data are generated from uniform distributions, the stochastic complexity test is quite efficient, and also the best of the three methods.

(iii). When the data are generated from lognormal distributions, the Smirnov test is the best and the other two tests are inefficient.

(iv). When the data are from exponential distributions, all the three methods are efficient, but the two sample t test is the best.

(v). When the data are from logistic distributions, both the Smirnov test and the stochastic complexity test are quite efficient to indicate a difference in the shape of the distributions with the latter method superior in performance.

(vi). When the data are from gamma distributions, both the Smirnov test and the stochastic complexity test are efficient with comparable power.

(vii). When the two samples are from different families of distributions, the stochastic complexity test is quite efficient and performs best.

From the simulation study it seems that the stochastic complexity test is a promising method which can be expected to be further improved as better ways to estimate the unknown densities are employed.

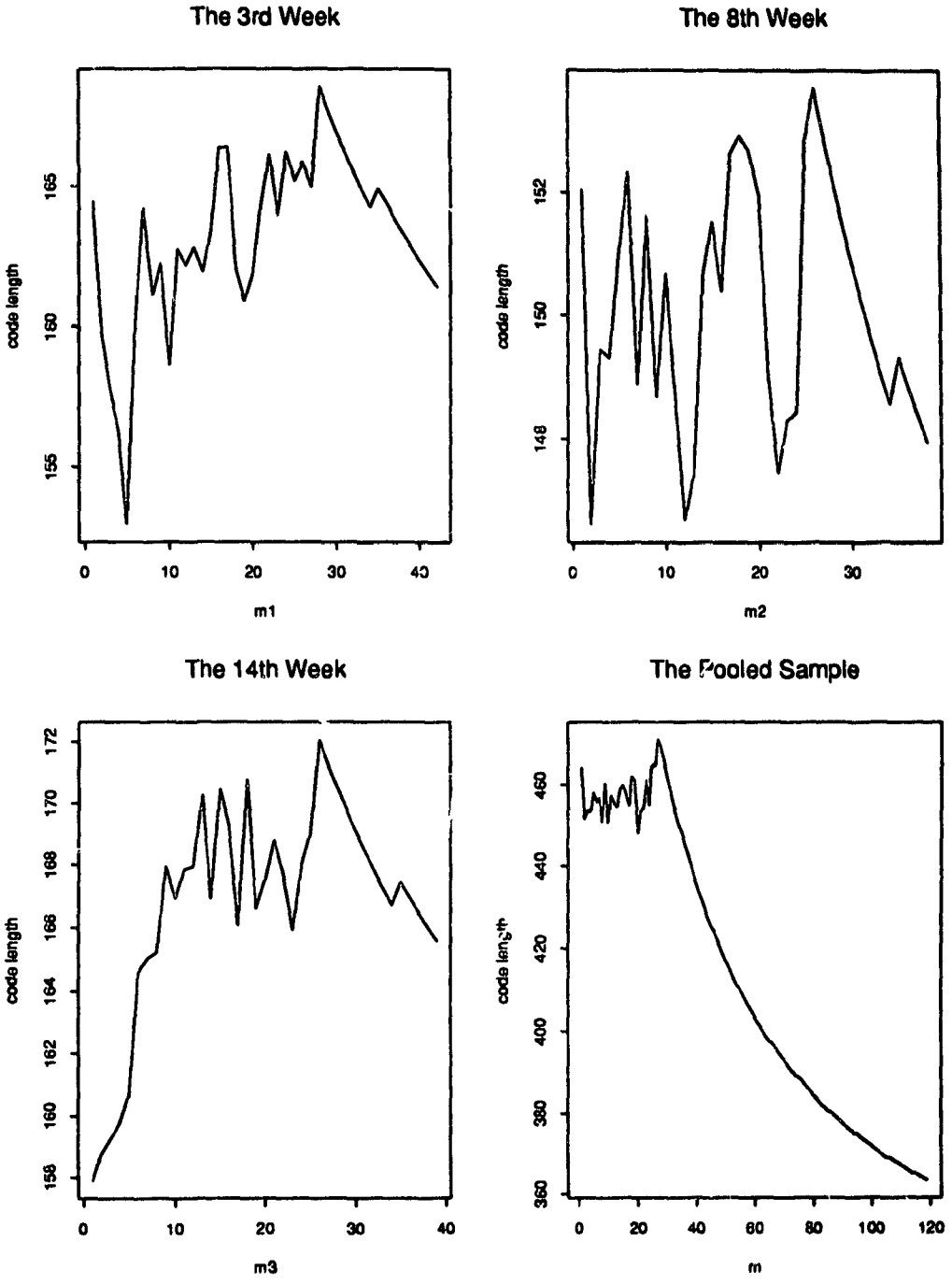


Figure 1: Relationship between Code Length and Number of Equal-width-bins

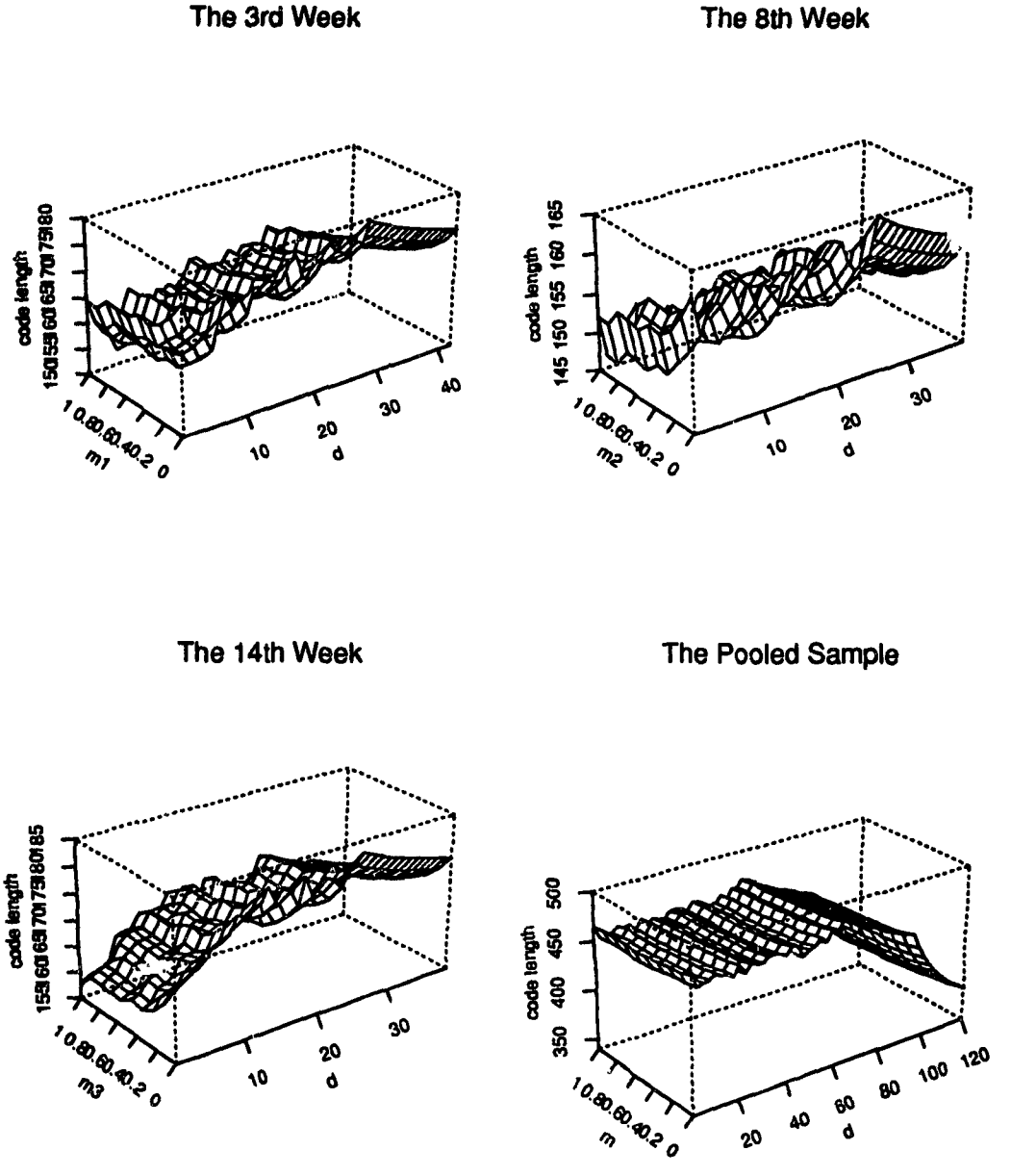


Figure 2: Relationship among Code Length, Number of Equal-width-bins and Precisions

Table 5.1: Comparison of the Power of the Test in Two Sample Case (Sizes $n_1 = n_2 = 15$ and Based on 1000 Simulations)

Distributions	Two Sample t-test			Smirnov Test			Stochastic Complexity Test						
	α: .1	.05	.01	α: .184	.076	.026	d: 0	.1	.2	.3	.5	.7	1.0
N(10, 9) & N(10, 49)	.107	.052	.011	.487	.229	.093	.410	.482	.547	.603	.783	.875	.970
N(10, 49) & N(10, 49)	.104	.048	.013	.173	.077	.023	.005	.012	.022	.032	.086	.173	.371
N(10, 9) & N(10, 9)	.094	.058	.012	.184	.078	.023	.033	.045	.079	.099	.220	.347	.447
N(12, 4) & N(10, 4)	.824	.717	.464	.837	.684	.503	.279	.333	.406	.454	.639	.748	.813
N(13, 4) & N(10, 4)	.988	.977	.894	.988	.952	.877	.655	.710	.758	.815	.891	.941	.965
N(13, 4) & N(13, 4)	.095	.044	.008	.185	.084	.025	.030	.047	.057	.079	.185	.326	.444
N(10, 4) & N(10, 4)	.108	.055	.007	.200	.083	.023	.033	.045	.073	.100	.235	.384	.427
N(10, 16) & N(10, 16)	.104	.049	.011	.201	.084	.031	.039	.056	.074	.104	.219	.362	.611
N(2, 16) & N(2, 16)	.090	.048	.015	.171	.078	.030	.017	.029	.033	.048	.113	.205	.513
N(10, 0.64) & N(10, 0.64)	.099	.049	.011	.186	.076	.020	.053	.073	.097	.133	.214	.396	.626
N(5, 0.64) & N(5, 0.09)	.095	.053	.008	.560	.311	.152	.873	.890	.902	.923	.986	.988	1.0
N(5, 0.09) & N(5, 0.09)	.111	.061	.007	.189	.077	.026	.137	.159	.231	.331	.687	.695	.964
N(25, 1) & N(25, 1)	.112	.056	.009	.201	.084	.032	.003	.005	.012	.018	.060	.131	.428
Unif(0, 1) & Unif(0, 1)	.094	.049	.011	.191	.087	.026	.103	.102	.203	.251	.844	.848	.851

Table 5.1 continued

Distributions	Two Sample t-test			Smirnov Test			Stochastic Complexity Test						
	α: .1	.05	.01	α: .184	.076	.026	d: 0	.1	.2	.3	.5	.7	1.0
Unif(1,4) & Unif(1,4)	.100	.052	.016	.182	.079	.029	.037	.039	.060	.055	.203	.470	.835
Unif(-2,3) & Unif(-2,3)	.087	.037	.007	.181	.068	.022	.011	.018	.021	.020	.145	.240	.851
Unif(1,4) & Unif(-2,3)	1.0	.998	.965	.998	.984	.937	.999	.999	.999	1.0	1.0	1.0	1.0
Unif(2,8) & Unif(3,7)	.100	.053	.009	.311	.124	.046	.244	.305	.425	.398	.807	.793	1.0
Unif(2,8) & Unif(2,8)	.086	.049	.007	.173	.071	.023	.007	.011	.015	.020	.079	.115	.465
Unif(3,7) & Unif(3,7)	.107	.061	.010	.206	.080	.030	.007	.011	.017	.028	.097	.096	.865
Unif(5,14) & Unif(3,10)	.957	.921	.762	.924	.814	.640	.909	.948	.970	.974	.990	1.0	.997
Unif(5,14) & Unif(5,14)	.103	.051	.008	.179	.070	.020	.001	.002	.004	.007	.033	.085	.090
LogN(1,1) & LogN(.5, $\sqrt{2}$)	.111	.050	.005	.410	.244	.129	.112	.129	.163	.193	.292	.399	.559
LogN(.5,$\sqrt{2}$) & LogN(.5, $\sqrt{2}$)	.070	.026	.003	.201	.082	.038	.108	.130	.158	.177	.233	.308	.418
LogN(1,1) & LogN(1,1)	.088	.035	.004	.165	.070	.023	.046	.062	.082	.112	.200	.291	.457
LogN(1,1) & LogN(0,$\sqrt{3}$)	.221	.151	.043	.737	.574	.379	.210	.250	.294	.338	.436	.521	.647
LogN(0,$\sqrt{3}$) & LogN(0,$\sqrt{3}$)	.055	.017	.000	.169	.077	.021	.182	.202	.225	.238	.296	.352	.439
LogN(7,2) & Log$N$(1,4)*	.491	.307	.080	1.0	1.0	.995	.510	.518	.536	.548	.568	.600	.628

*: The power will equal to .802, .875 and .925 respectively for d=3.0,4.0 and 5.0.

Table 5.1 continued

Distributions	Two Sample t-test			Smirnov Test			Stochastic Complexity Test						
	α: .1	.05	.01	α: .184	.076	.026	d: 0	.1	.2	.3	.5	.7	1.0
Exp(1) & Exp(.2)	.989	.939	.632	.982	.946	.848	.917	.940	.958	.961	.982	.992	.994
Exp(.2) & Exp(.5)	.719	.545	.193	.716	.525	.359	.416	.462	.509	.544	.661	.769	.841
Exp(.2) & Exp(.6)	.862	.721	.336	.854	.705	.526	.599	.646	.683	.729	.824	.886	.923
Exp(.2) & Exp(.2)	.100	.049	.007	.181	.077	.021	.017	.026	.039	.053	.129	.216	.370
Exp(.2) & Exp(.7)	.919	.823	.449	.898	.798	.651	.728	.765	.804	.829	.889	.938	.960
Exp(.6) & Exp(.6)	.089	.043	.005	.185	.088	.026	.089	.108	.133	.164	.338	.411	.682
Exp(.7) & Exp(.7)	.095	.039	.006	.179	.082	.017	.110	.134	.162	.192	.342	.428	.745
Logis(2,2) & Logis(2,2)	.093	.046	.007	.190	.075	.023	.026	.038	.057	.084	.175	.298	.584
Logis(2,3) & Logis(2,4)	.096	.047	.008	.209	.092	.034	.014	.021	.031	.048	.093	.173	.345
Logis(2,3) & Logis(2,3)	.112	.056	.007	.177	.079	.025	.008	.010	.020	.031	.078	.155	.338
Logis(2,5) & Logis(2,5)	.107	.052	.013	.180	.076	.033	.002	.003	.004	.011	.032	.060	.151
Logis(2,3) & Logis(2,5)	.079	.034	.010	.266	.118	.046	.032	.049	.064	.087	.157	.257	.425
Logis(2,2) & Logis(2,7)	.088	.047	.006	.714	.409	.192	.517	.580	.643	.710	.817	.879	.938
Logis(2,7) & Logis(2,7)	.104	.054	.005	.189	.075	.024	.002	.003	.004	.008	.024	.044	.116

Table 5.1 continued

Distributions	Two Sample t-test			Smirnov Test			Stochastic Complexity Test						
	α: .1	.05	.01	α: .184	.076	.026	d: 0	.1	.2	.3	.5	.7	1.0
Logis(2,3) & Logis(2,7)	.100	.045	.011	.460	.224	.084	.132	.169	.210	.267	.387	.518	.697
Logis(2,3) & Logis(2,8)	.113	.055	.010	.573	.300	.135	.238	.270	.327	.379	.489	.623	.765
Logis(2,5) & Logis(2,8)	.113	.058	.010	.266	.108	.049	.013	.019	.028	.039	.092	.156	.278
Logis(2,8) & Logis(2,8)	.103	.047	.008	.186	.070	.022	.002	.003	.004	.004	.009	.024	.078
Logis(2,4) & Logis(2,7)	.092	.042	.005	.307	.136	.042	.033	.043	.052	.070	.138	.247	.419
Gamma(4,2) & Gamma(2,3)	.392	.271	.102	.566	.370	.210	.089	.113	.153	.196	.328	.471	.712
Gamma(2,3) & Gamma(2,3)	.099	.050	.007	.180	.080	.029	.030	.040	.055	.066	.154	.250	.457
Gamma(4,2) & Gamma(4,2)	.096	.047	.010	.167	.069	.024	.019	.037	.059	.080	.166	.245	.538
Gamma(2,4) & Gamma(4,2)	.108	.057	.018	.260	.115	.044	.048	.071	.091	.121	.244	.364	.639
Gamma(5,2) & Gamma(2,5)	.116	.050	.012	.318	.150	.062	.069	.097	.131	.171	.308	.422	.644
Gamma(2,5) & Gamma(2,5)	.100	.050	.011	.187	.085	.025	.012	.020	.029	.036	.079	.137	.289
Gamma(5,3) & Gamma(3,5)	.101	.049	.006	.226	.119	.041	.012	.018	.031	.046	.106	.187	.305
Gamma(5,1) & Gamma(1,5)	.132	.088	.040	.587	.368	.200	.384	.434	.488	.547	.715	.825	.917
Gamma(6,2) & Gamma(2,6)	.109	.056	.014	.370	.172	.084	.085	.104	.155	.187	.327	.471	.624

Table 5.1 continued

Distributions	Two Sample t-test			Smirnov Test			Stochastic Complexity Test						
	α: .1	.05	.01	α: .184	.076	.026	d: 0	.1	.2	.3	.5	.7	1.0
Gamma(7,2) & Gamma(2,7)	.114	.067	.005	.390	.216	.092	.092	.136	.169	.224	.366	.519	.635
Gamma(7,3) & Gamma(3,7)	.110	.062	.019	.292	.155	.067	.028	.039	.054	.073	.145	.239	.447
Gamma(8,2) & Gamma(2,8)	.094	.041	.009	.455	.242	.111	.119	.157	.206	.259	.427	.566	.707
Gamma(8,3) & Gamma(3,8)	.114	.063	.013	.315	.158	.065	.025	.038	.060	.087	.156	.258	.437
Gamma(9,2) & Gamma(2,9)	.117	.067	.019	.509	.286	.131	.158	.209	.246	.310	.463	.609	.745
Gamma(5,2) & Gamma(5,2)	.088	.036	.003	.166	.065	.026	.016	.025	.046	.065	.150	.236	.497
Gamma(5,3) & Gamma(5,3)	.105	.054	.014	.207	.082	.028	.008	.010	.018	.026	.088	.152	.247
Gamma(3,5) & Gamma(3,5)	.091	.050	.013	.177	.074	.029	.006	.009	.012	.018	.052	.109	.236
Gamma(5,1) & Gamma(5,1)	.109	.058	.009	.186	.070	.025	.071	.094	.135	.175	.304	.513	.590
Gamma(1,5) & Gamma(1,5)	.099	.049	.012	.189	.089	.036	.021	.032	.047	.061	.121	.213	.372
Gamma(6,2) & Gamma(6,2)	.095	.049	.008	.190	.079	.022	.016	.021	.036	.062	.149	.246	.398
Gamma(2,6) & Gamma(2,6)	.108	.054	.007	.202	.086	.038	.004	.007	.012	.020	.058	.121	.260
Gamma(7,2) & Gamma(7,2)	.067	.030	.009	.138	.044	.016	.012	.018	.025	.035	.097	.175	.315
Gamma(2,7) & Gamma(2,7)	.098	.039	.005	.185	.074	.020	.004	.005	.009	.015	.049	.101	.222

Table 5.1 continued

Distributions	Two Sample t-test			Smirnov Test			Stochastic Complexity Test						
	α: .1	.05	.01	α: .184	.076	.026	d: 0	.1	.2	.3	.5	.7	1.0
Gamma(7,3) & Gamma(7,3)	.118	.063	.010	.199	.092	.029	.002	.006	.014	.026	.072	.132	.281
Gamma(3,7) & Gamma(3,7)	.102	.054	.006	.178	.072	.022	.004	.006	.012	.015	.036	.075	.170
Gamma(8,2) & Gamma(8,2)	.081	.048	.010	.183	.083	.026	.008	.022	.026	.041	.092	.168	.299
Gamma(2,8) & Gamma(2,8)	.095	.046	.007	.181	.083	.028	.006	.009	.009	.013	.036	.080	.207
Gamma(8,3) & Gamma(8,3)	.115	.057	.008	.178	.074	.024	.002	.004	.007	.008	.036	.087	.231
Gamma(3,8) & Gamma(3,8)	.100	.049	.014	.190	.073	.030	.002	.006	.009	.011	.022	.051	.150
Gamma(9,2) & Gamma(9,2)	.113	.065	.015	.192	.087	.032	.008	.008	.014	.020	.068	.151	.295
Gamma(2,9) & Gamma(2,9)	.097	.047	.010	.188	.078	.033	.001	.002	.003	.008	.024	.066	.158
N(2,16) & Logis(2,7)	.118	.060	.004	.629	.367	.166	.409	.490	.562	.635	.748	.845	.939
N(5,5) & Exp(.2)	.125	.075	.027	.562	.335	.173	.359	.421	.487	.544	.758	.841	.916
N(2,49π^2/3) & Logis(2,7)	.101	.056	.005	.202	.087	.024	.000	.001	.003	.004	.012	.030	.087
N(5,36) & Exp(.2)	.090	.048	.009	.323	.156	.065	.197	.240	.287	.348	.537	.692	.849

Table 5.2: Comparison of the Power of the Test in Two Sample Case (Sizes $n_1 = 15, n_2 = 20$ and Based on 1000 Simulations)

Distributions	Two Sample t-test			Smirnov Test				Stochastic Complexity Test						
	α: .1	.05	.01	α: .2	.1	.05	.01	d: 0	.1	.2	.3	.5	.7	1.0
Unif(-2,3) & Unif(1,4)	1.0	.997	.973	.998	.995	.984	.917	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Unif(3,7) & Unif(1,4)	1.0	1.0	1.0	1.0	1.0	1.0	.999	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Unif(2,8) & Unif(3,7)	.103	.051	.011	.362	.239	.125	.029	.505	.593	.684	.666	.914	.897	1.0
Unif(2,8) & Unif(4,7)	.232	.146	.050	.655	.555	.329	.149	.966	.974	.974	.990	.997	.997	1.0
Unif(2,8) & Unif(1,9)	.102	.054	.012	.274	.139	.083	.016	.111	.156	.235	.295	.405	.811	.858
Unif(1,9) & Unif(1,9)	.108	.045	.015	.224	.117	.068	.020	.009	.017	.024	.027	.077	.318	.290
Unif(4,7) & Unif(4,7)	.102	.045	.011	.200	.104	.056	.011	.021	.026	.039	.047	.104	.225	.863
Unif(-2,3) & Unif(-3,4)	.079	.040	.007	.287	.155	.084	.014	.093	.140	.165	.182	.454	.885	1.0
Unif(-3,4) & Unif(-3,4)	.094	.048	.012	.203	.105	.057	.013	.006	.009	.012	.011	.033	.159	.295
Unif(-3,4) & Unif(-4,0)	.998	.997	.974	.996	.992	.961	.849	.991	.994	.995	.997	1.0	1.0	1.0
Unif(-4,0) & Unif(-4,0)	.103	.054	.011	.200	.088	.053	.011	.005	.007	.006	.014	.037	.037	.865
LogN(1,1) & LogN(.5,$\sqrt{2}$)	.092	.035	.004	.470	.317	.237	.079	.109	.134	.177	.211	.310	.403	.546
Exp(.2) & Exp(.6)	.873	.747	.350	.887	.797	.703	.475	.635	.680	.729	.761	.837	.892	.925
Exp(.3) & Exp(.8)	.857	.759	.432	.808	.698	.598	.371	.585	.634	.685	.723	.809	.885	.916

Table 5.2 continued

Distributions	*Two Sample t-test*			*Smirnov Test*				Stochastic Complexity Test						
	α: .1	.05	.01	α: .2	.1	.05	.01	d: 0	.1	.2	.3	.5	.7	1.0
Exp(.5) & Exp(.2)	.771	.602	.238	.780	.647	.548	.284	.403	.454	.522	.569	.673	.768	.845
Logis(2,2) & Logis(2,7)	.085	.044	.012	.774	.543	.359	.097	.559	.616	.678	.733	.833	.910	.955
Logis(2,4) & Logis(2,7)	.061	.030	.008	.322	.169	.099	.018	.028	.042	.053	.073	.144	.233	.390
Logis(2,8) & Logis(2,3)	.094	.040	.008	.623	.430	.264	.090	.303	.342	.388	.434	.555	.669	.794
N(10,9) & N(10,49)	.097	.046	.009	.553	.350	.221	.047	.478	.540	.620	.679	.828	.912	.976
N(13,4) & N(10,4)	.998	.992	.941	.998	.986	.980	.881	.756	.796	.852	.887	.937	.964	.982
N(5,.64) & N(5,.09)	.098	.049	.011	.674	.482	.285	.092	.926	.926	.934	.956	.994	.994	.999
N(-8,49) & N(-6,16)	.325	.204	.074	.542	.383	.263	.110	.101	.129	.171	.220	.341	.473	.625
Gamma(7,2) & Gamma(2,7)	.089	.054	.013	.458	.274	.192	.066	.134	.171	.218	.269	.433	.591	.723
Gamma(7,2) & Gamma(8,3)	.993	.978	.903	.993	.972	.953	.843	.575	.636	.677	.736	.827	.898	954
Gamma(4,3) & Unif(6,18)	.120	.066	.017	.323	.179	.100	.020	.391	.449	.536	.597	.752	.885	.922
N(3,16) & Logis(2,7)	.089	.045	.004	.727	.499	.325	.090	.475	.545	.617	.674	.803	.872	.948
N(5,25) & Exp(.25)	.202	.121	.040	.514	.358	.263	.102	.363	.422	.489	.555	.743	.834	.924
Gamma(5,2) & N(11,81)	.099	.049	.012	.528	.315	.200	.060	.290	.346	.414	.461	.649	.802	.900

5.4 Proofs of the Theorems

In this section, we will provide proofs for all the theorems listed in this chapter. For the sake of simplicity, the logarithms in the proofs are al' natural logarithms.

From (5.2.6) and (5.2.14),

$$-\log \tilde{f}(X^n;m) + L_1^*(X^n;m) =$$
$$-\log\left\{\frac{(m-1)!}{(n+m-1)!}\prod_{i=1}^{m}\frac{n_{i,m}!}{\tilde{r}_{i,m}^{\,n_{i,m}}}\right\} + \sum_{i=1}^{m}(n_{i,m}+1)\log\frac{n_{i,m}+1}{(n+m)\tilde{r}_{i,m}}. \qquad (5.4.1)$$

By Stirling's formula $n! = \sqrt{2\pi n}n^n e^{-n}e^{\theta_n}$ $(0 < \theta_n < 1/(12n))$, (5.4.1) can be rewritten as

$$-\log \tilde{f}(X^n;m) + L_1^*(X^n;m) = -\sum_{i=1}^{m}\log \tilde{r}_{i,m} + \sum_{n_{i,m}>0}\log\left(1+\frac{1}{n_{i,m}}\right)^{n_{i,m}+1}$$
$$+\frac{1}{2}\sum_{n_{i,m}>0}\log n_{i,m} - m\log m + O(m). \qquad (5.4.2)$$

We will show that the first term of (5.4.2) is $\alpha_{12}m\log m + O(m)$ where $1 \le \alpha_{12} \le \alpha_1$, the second term is $O(m)$ and the third term is $\frac{m}{2}\log\frac{n}{m^{\alpha_{22}}}$ where $1 \le \alpha_{22} \le \alpha_1$. The following Lemmas will be needed.

Lemma 5.4.1 *Suppose that $n_{i,m}$'s have a multinomial distribution with probabilities $\pi_{i,m}$'s such that $\sum_{i=1}^{m}\pi_{i,m} = 1$, $\pi_{i,m} \ge b_1 c_1 i^{-\alpha_1}$ and $\sum_{i=1}^{m} n_{i,m} = n$. Then for each integer w, there exists a constant a_w such that*

$$E\left\{\sum_{i=1}^{m}\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right\}^{2w} \le a_w n^{\frac{1}{2}-w}m^{2\alpha_1 w} \qquad (5.4.3)$$

Proof. Denote $T_1 = \sum_{i=1}^{m}\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}$. By the definition of the multinomial distribution and from Stirling's formula we have

$$E\left(T_1^{2w}\right) = \sum_{n_{1,m}+\cdots+n_{m,m}=n} T_1^{2w}\frac{n!}{\prod_{i=1}^{m} n_{i,m}!}\prod_{i=1}^{m}\pi_{i,m}^{\,n_{i,m}}$$

$$
\begin{aligned}
&= \sum_{n_{1,m}+\cdots+n_{m,m}=n} T_1^{2w} \prod_{i=1}^{m} \frac{(n\pi_{i,m})^{n_{i,m}}}{n_{i,m}!} e^{-n_{i,m}} \sqrt{2\pi n} e^{c_n} \\
&\leq \sqrt{2\pi n} e \sum_{N=0}^{\infty} \sum_{n_{1,m}+\cdots+n_{m,m}=N} T_1^{2w} \prod_{i=1}^{m} \frac{(n\pi_{i,m})^{n_{i,m}}}{n_{i,m}!} e^{-n_{i,m}} \\
&= \sqrt{2\pi n} e \sum_{n_{1,m}=0}^{\infty} \cdots \sum_{n_{m,m}=0}^{\infty} T_1^{2w} \prod_{i=1}^{m} \frac{(n\pi_{i,m})^{n_{i,m}}}{n_{i,m}!} e^{-n_{i,m}} \\
&= \sqrt{2\pi n} e E'\left(T_1^{2w}\right) \qquad (5.4.4)
\end{aligned}
$$

where the final expectation $E'(T_1^{2w})$ is with respect to a series of independent Poisson random variables $\{n_{i,m}\}$ with parameters $\{n\pi_{i,m}\}$. This technique, used by Rosenblatt (1975) and Stone (1985), of converting the multinomial to Poisson is called Poissonization. The constant $c_n = o\left(\frac{1}{12n}\right)$. By Shiryayev's (1984) Theorem 6 of Section 2.12, the $2w$-th moments of T_1 can be written as a sum of its cumulants:

$$
E'\left(T_1^{2w}\right) = \sum_{j_1+\cdots+j_l=2w} \rho(j_1,\cdots,j_l) \prod_{k=1}^{l} \kappa_{j_k}(T_1) \qquad (5.4.5)
$$

where $\rho(j_1,\cdots,j_l) = \frac{1}{l!}\frac{(2w)!}{j_1!\cdots j_l!}$ and $j_k \geq 1, l \leq 2w$. Because $n_{i,m}$'s are independent Poisson random variables, it follows from the section 1.4 of Lehmann (1986a) that the j_k-th cumulants of T_1

$$
\kappa_{j_k}(T_1) = \sum_{i=1}^{m} \kappa_{j_k}\left(\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right) = \sum_{i=1}^{m} \frac{n\pi_{i,m}}{(n\pi_{i,m})^{j_k}} \leq (b_1c_1)^{-j_k} n^{1-j_k} m^{\alpha_1 j_k} \qquad (5.4.6)
$$

if $j_k > 1$ and $\kappa_{j_k}(T_1) = 0$ if $j_k = 1$. Thus

$$
\begin{aligned}
E'\left(T_1^{2w}\right) &= \sum{}^{*} \rho(j_1,\cdots,j_l) \prod_{k=1}^{l} \kappa_{j_k}(T_1) \\
&\leq \sum{}^{*} \rho(j_1,\cdots,j_l) \prod_{k=1}^{l} (b_1c_1)^{-j_k} n^{1-j_k} m^{\alpha_1 j_k} \leq a_w n^{-w} m^{2\alpha_1 w} \qquad (5.4.7)
\end{aligned}
$$

where the summation $\sum^{*}$ is taken over all partitions of $2w$ such that $\sum_{k=1}^{l} j_k = 2w$, $j_k \geq 2$ and $l \leq w$. Using the same notation for possibly different constants and substituting the last bound into (5.4.4) the lemma is proved. □

Lemma 5.4.2 *Suppose that N is a binomial random variable with mean np. Then for any integer $w > 0$, there is a constant $a_w > 0$ such that*

$$E(N - np)^{2w} \leq a_w n^{\frac{1}{2}+w} p^w \tag{5.4.8}$$

Proof. By the same technique of Poissonization we have

$$E(N - np)^{2w} \leq a_w n^{\frac{1}{2}} E(N_1 - np)^{2w}$$

where N_1 is a Poisson random variable with mean np. By the equation (5.4.5) and the fact that $\kappa_k(N_1 - np) = np$ if $k > 1$ and $\kappa_1(N_1 - np) = 0$ it follows that $E(N_1 - np)^{2w}$ is a polynomial of order w, and therefore (5.4.8) must hold. □

Lemma 5.4.3 *Under the conditions that $\tilde{r}_{i,m} \geq b_1 m^{-\alpha_1}$, $1 \leq \alpha_1 < 1 + \frac{1}{2\gamma_2}$ and $f \geq c_1 > 0$,*

$$\sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} = o(m) \quad a.s. \tag{5.4.9}$$

uniformly in $m \in [1, n^{\gamma_2}]$ as $n \to \infty$, where $\pi_{i,m} = \int_{\tilde{Q}_{i,m}} f$

Proof. For any $\varepsilon > 0$,

$$\begin{aligned} & P\left(\max_{m \in [1, n^{\gamma_2}]} \left| \sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right| \geq \varepsilon m \right) \\ & \leq \sum_{m \in [1, n^{\gamma_2}]} P\left(\left| \sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right| \geq \varepsilon m \right) \\ & \leq \sum_{m \in [1, n^{\gamma_2}]} \varepsilon^{-2w} m^{-2w} E\left\{ \sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right\}^{2w} \end{aligned} \tag{5.4.10}$$

where the last inequality is obtained by applying Chebyshev's inequality. From Lemma 5.4.1,

$$\begin{aligned} P\left(\max_{m \in [1, n^{\gamma_2}]} \left| \sum_{i=1}^{m} \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} \right| \geq \varepsilon m \right) & \leq \sum_{m \in [1, n^{\gamma_2}]} \varepsilon^{-2w} m^{-2w} a_w n^{\frac{1}{2}-w} m^{2\alpha_1 w} \\ & \leq a_w \varepsilon^{-2w} n^{(2\alpha_1\gamma_2 - 2\gamma_2 - 1)w + \frac{1}{2} + \gamma_2} \end{aligned} \tag{5.4.11}$$

By the condition that $\alpha_1 < 1 + \frac{1}{2\gamma_2}$, the series above converges in n for $w > \frac{3+2\gamma_2}{2+4\gamma_2-4\alpha_1\gamma_2}$. Hence (5.4.9) follows from the Borel-Cantelli lemma. □

Lemma 5.4.4 *Under the conditions that* $b_1 m^{-\alpha_1} \le \tilde{r}_{i,m} \le b_2 m^{-\alpha_2}$, *where* $2\alpha_1 - \frac{1}{\gamma_2} < \alpha_2 \le 1 \le \alpha_1 < \frac{1}{2} + \frac{1}{2\gamma_2}$ *and* $0 < c_1 \le f \le c_2$,

$$\max_{1\le i\le m} \left| \frac{m^{\alpha_1} n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m} \right| = o(1) \quad a.s. \tag{5.4.12}$$

uniformly in $m \in [1, n^{\gamma_2}]$ *as* $n \to \infty$.

Proof. Denote

$$I_{m,n} = \max_{1\le i\le m} \left| \frac{m^{\alpha_1} n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m} \right|,$$

then for any $\varepsilon > 0$,

$$\begin{aligned} P\left(\max_{m\in[1,n^{\gamma_2}]} I_{m,n} > \varepsilon \right) &\le \sum_{m\in[1,n^{\gamma_2}]} P\left(I_{m,n} > \varepsilon\right) \\ &\le \sum_{m\in[1,n^{\gamma_2}]} \sum_{i=1}^{m} P\left(\left| \frac{m^{\alpha_1} n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m} \right| > \varepsilon \right) \\ &\le \sum_{m\in[1,n^{\gamma_2}]} \sum_{i=1}^{m} \varepsilon^{-2w} E \left| \frac{m^{\alpha_1} n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m} \right|^{2w} \\ &= \sum_{m\in[1,n^{\gamma_2}]} \sum_{i=1}^{m} \varepsilon^{-2w} m^{2w\alpha_1} n^{-2w} E(n_{i,m} - n\pi_{i,m})^{2w}, \end{aligned} \tag{5.4.13}$$

where the last inequality is obtained by applying Chebyshev's inequality. From Lemma 5.4.2 and the property that $c_1 b_1 m^{-\alpha_1} \le \pi_{i,m} \le c_2 b_2 m^{-\alpha_2}$,

$$\begin{aligned} P\left(\max_{m\in[1,n^{\gamma_2}]} I_{m,n} > \varepsilon \right) &\le \sum_{m\in[1,n^{\gamma_2}]} \sum_{i=1}^{m} \varepsilon^{-2w} m^{2w\alpha_1} n^{-2w} a_w n^{\frac{1}{2}+w} (c_2 b_2 m^{-\alpha_2})^w \\ &\le a_w n^{2\gamma_2 + \frac{1}{2} + (2\alpha_1\gamma_2 - \alpha_2\gamma_2 - 1)w}. \end{aligned} \tag{5.4.14}$$

From now on the same notation will be used for possibly different constants. By the condition that $\alpha_2 > 2\alpha_1 - \frac{1}{\gamma_2}$, the above series converges in n for $w > \frac{4\gamma_2+3}{2+2\alpha_2\gamma_2-4\alpha_1\gamma_2}$. Hence (5.4.12) follows again from the Borel-Cantelli Lemma. □

Lemma 5.4.5 *Under the conditions that* $\tilde{r}_{i,m} \le b_2 m^{-\alpha_2}$ *and* $f \le c_2$, *we have*

$$\sum_{i=1}^{m} (n_{i,m} - n\pi_{i,m})^2 - n = o(n) \quad a.s. \tag{5.4.15}$$

uniformly in $m \in [n^{\gamma_1}, n]$ *as* $n \to \infty$.

Proof. Suppose $\{N_{i,m}\}$ are a sequence of independent Poisson random variables with mean $\{n\pi_{i,m}\}$ and denote $T_2 = \sum_{i=1}^{m}(N_{i,m} - n\pi_{i,m})^2$. We first show that the j-th cumulants of T_2 satisfies

$$|\kappa_j(T_2)| \leq a_j n^j m^{-\alpha_2(j-1)} \tag{5.4.16}$$

where a_j is a constant depending on j.

Because $\{N_{i,m}\}$ are independent, it follows that

$$\kappa_j(T_2) = \sum_{i=1}^{m} \kappa_j \left((N_{i,m} - n\pi_{i,m})^2\right).$$

By applying again Theorem 6 of Section 2.12 of Shiryayev's (1984), the j-th cumulants of $(N_{i,m} - n\pi_{i,m})^2$ can be written as a sum of its moments:

$$\kappa_j \left((N_{i,m} - n\pi_{i,m})^2\right) = \sum_{j_1+\cdots+j_l=j} \zeta(j_1,\cdots,j_l) \prod_{k=1}^{l} E\left((N_{i,m} - n\pi_{i,m})^{2j_k}\right) \tag{5.4.17}$$

where $\zeta(j_1,\cdots,j_l) = \frac{(-1)^{l-1}}{l}\frac{j!}{j_1!\cdots j_l!}$ and $j_k \geq 1$, $l \leq j$. From Lemma 5.4.2 we know that $E\left((N_{i,m} - n\pi_{i,m})^{2j_k}\right)$ is a polynomial of order j_k for $n\pi_{i,m}$, therefore

$$|\kappa_j(T_2)| \leq \sum_{i=1}^{m} a_j (n\pi_{i,m})^j \leq a_j n^j m^{-\alpha_2(j-1)}$$

for some constant a_j, hence (5.4.16) holds.

By (5.4.5) and the identities $\kappa_1(T_2 - n) = E(T_2 - n) = 0$ and $\kappa_j(T_2 - n) = \kappa_j(T_2)$ for $j \geq 2$, it can be seen that

$$E(T_2 - n)^{2w} = \sum{}^{*} \rho(l_1,\cdots,l_k) \prod_{j=1}^{k} \kappa_{l_j}(T_2)$$

where the summation $\sum^{*}$ is taken over all the partitions of $2w$ such that $\sum_{j=1}^{k} l_j = 2w$, $l_j \geq 2$ and $k \leq w$. By (5.4.16) it follows that

$$E(T_2 - n)^{2w} \leq \sum{}^{*} a_w n^{2w} m^{-\alpha_2(2w-k)} \leq a_w n^{2w} m^{-\alpha_2 w} \tag{5.4.18}$$

for some constant a_w depending on w.

Now for any $\varepsilon > 0$,

$$\begin{aligned} & P\left(\max_{m\in[n^{\gamma_1},n]}\left|\sum_{i=1}^{m}(n_{i,m}-n\pi_{i,m})^2-n\right|>\varepsilon n\right) \\ \leq & \sum_{m\in[n^{\gamma_1},n]} P\left(\left|\sum_{i=1}^{m}(n_{i,m}-n\pi_{i,m})^2-n\right|>\varepsilon n\right) \\ \leq & \sum_{m\in[n^{\gamma_1},n]} \varepsilon^{-2w}n^{-2w}E\left|\sum_{i=1}^{m}(n_{i,m}-n\pi_{i,m})^2-n\right|^{2w} \\ \leq & \sum_{m\in[n^{\gamma_1},n]} \varepsilon^{-2w}n^{-2w+\frac{1}{2}}E\left|\sum_{i=1}^{m}(N_{i,m}-n\pi_{i,m})^2-n\right|^{2w} \end{aligned} \tag{5.4.19}$$

by applying Chebyshev's inequality and the technique of Poissonization.

From (5.4.18) it follows that

$$\begin{aligned} & P\left(\max_{m\in[n^{\gamma_1},n]}\left|\sum_{i=1}^{m}(n_{i,m}-n\pi_{i,m})^2-n\right|>\varepsilon n\right) \\ \leq & \sum_{m\in[n^{\gamma_1},n]} \varepsilon^{-2w}n^{-2w+\frac{1}{2}}a_w n^{2w}m^{-\alpha_2 w} \leq a_w n^{\frac{3}{2}-\alpha_2\gamma_1 w}. \end{aligned} \tag{5.4.20}$$

The above series converges for $w > \frac{5}{2\alpha_2\gamma_1}$, hence from the Borel-Cantelli Lemma (5.4.15) follows. □

Corollary 5.4.1 *Under the conditions that* $b_1 m^{-\alpha_1} \leq \tilde{r}_{i,m} \leq b_2 m^{-\alpha_2}$ *and* $0 < c_1 \leq f \leq c_2$,

$$\sum_{i=1}^{m}\frac{(n_{i,m}-n\pi_{i,m})^2}{(n\pi_{i,m})^2} = O\left(\frac{m^{2\alpha_1}}{n}\right) \quad a.s. \tag{5.4.21}$$

uniformly in $m \in [n^{\gamma_1}, n]$ *as* $n \to \infty$.

Lemma 5.4.6 *Under the conditions of Lemma 5.4.4, the following statement is true:*

$$\sum_{n_{i,m}>0}\log\frac{n_{i,m}}{n\pi_{i,m}} = O(m) \quad a.s. \tag{5.4.22}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ *as* $n \to \infty$.

Proof. First note that

$$\sum_{n_{i,m}>0}\log\frac{n_{i,m}}{n\pi_{i,m}} = \sum_{n_{i,m}>0}\log\left(1+\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right)$$

By Taylor expansion,

$$\sum_{n_{i,m}>0} \log \frac{n_{i,m}}{n\pi_{i,m}} = \sum_{i=1}^{m} \frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}} - \frac{1}{2} \sum_{n_{i,m}>0} (1+\xi_{i,m})^{-2} \frac{(n_{i,m}-n\pi_{i,m})^2}{(n\pi_{i,m})^2} + D \quad (5.4.23)$$

where $D = -\sum_{n_{i,m}=0} \frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}} < m$ and $|\xi_{i,m}| \le \left|\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right|$.

Thus

$$\max_{1\le i\le m} |\xi_{i,m}| \le \max_{1\le i\le m} \left|\frac{n_{i,m}-n\pi_{i,m}}{n\pi_{i,m}}\right| \le (c_1 b_1)^{-1} \max_{1\le i\le m} \left|\frac{m^{\alpha_1} n_{i,m}}{n} - m^{\alpha_1}\pi_{i,m}\right|,$$

and by Lemma 5.4.4

$$\max_{1\le i\le m} |\xi_{i,m}| = o(1) \quad \text{a.s.} \quad (5.4.24)$$

uniformly in $m \in [1, n^{\gamma_2}]$ as $n \to \infty$. By (5.4.24) and Corollary 5.4.1 it follows that the second term of the right hand side of (5.4.23) is bounded uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ by $O\left(\frac{m^{2\alpha_1}}{n}\right)$ a.s.. The latter is $o(m)$ because $n > m^{\frac{1}{\gamma_2}}$ and $\alpha_1 < \frac{1}{2} + \frac{1}{2\gamma_2}$. Therefore by Lemma 5.4.3

$$\sum_{n_{i,m}>0} \log \frac{n_{i,m}}{n\pi_{i,m}} = O(m) \quad \text{a.s.}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$. □

Proof of Theorem 5.2.1

By conditions (i) and (iii) we can obtain an interval estimate, respectively, for $-\sum_{i=1}^{m} \tilde{r}_{i,m}$ and $\sum_{i=1}^{m} \log n\pi_{i,m}$ as follows:

$$m\log m + O(m) \le -\sum_{i=1}^{m} \log \tilde{r}_{i,m} \le \alpha_1 m \log m + O(m) \quad (5.4.25)$$

$$m\log n - \alpha_1 m\log m + O(m) \le \sum_{i=1}^{m} \log n\pi_{i,m} \le m\log n - m\log m + O(m). \quad (5.4.26)$$

Hence there exists an α' satisfying $-\frac{1}{2}\alpha_1 \le \alpha' \le -\frac{3}{2} + \alpha_1$ such that

$$-\sum_{i=1}^{m} \log \tilde{r}_{i,m} + \frac{1}{2}\sum_{i=1}^{m} \log n\pi_{i,m} - m\log m = \alpha' m\log m + \frac{1}{2} m\log n + O(m). \quad (5.4.27)$$

Now we turn to the second term of (5.4.2). By Taylor expansion

$$\sum_{n_{i,m}>0} \log \left(1+\frac{1}{n_{i,m}}\right)^{n_{i,m}+1}$$

$$= \sum_{n_{i,m}>0} (n_{i,m}+1)\left(\frac{1}{n_{i,m}} - \frac{1}{2}(1+\eta_{i,m})^{-2}\frac{1}{{n_{i,m}}^2}\right) = O(m), \quad (5.4.28)$$

where $0 \le \eta_{i,m} \le \frac{1}{n_{i,m}}$.

From Lemma 5.4.6, (5.4.27), (5.4.28) and (5.4.2), it is easy to see that

$$-\log \tilde{f}(X^n; m) + L_1^*(X^n; m) = \alpha' m \log m + \frac{1}{2} m \log n + O(m) \quad \text{a.s.}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$. □

To prove Theorem 5.2.2 we first need the following lemmas.

Lemma 5.4.7 *Under the condition (iii) of Theorem 5.2.1,*

$$L_2(\tilde{q}^m, m, \delta) = o(m) \tag{5.4.29}$$

Proof. From $b_1 m^{-\alpha_1} \le \tilde{r}_{i,m} \le b_2 m^{-\alpha_2}$ it follows that

$$\left| \overline{\tilde{r}_{i,m} - \frac{r}{m}} \right| \le \max\left\{ \frac{b_2}{m^{\alpha_2}} - \frac{r}{m}, \frac{r}{m} - \frac{b_1}{m^{\alpha_1}} \right\} \le \frac{b_2 + r}{m^{\alpha_2}}.$$

From this (5.4.29) follows. □

Let $f(x \mid \tilde{q}^m)$ denote a density in H_m which assigns the same probability as f to each subinterval $\tilde{Q}_{i,m}$, i.e. for $x \in [s, t]$ let

$$f(x \mid \tilde{q}^m) = \sum_{i=1}^{m} \frac{\pi_{i,m}}{\tilde{r}_{i,m}} I_{\tilde{Q}_{i,m}}(x).$$

By Lemma 5.4.7 we have

$$\begin{aligned} &-L_1^*(X^n; m) + L_2(\tilde{q}^m, m, \delta) + \log f^n(X^n) \\ &= -L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j \mid \tilde{q}^m) + \sum_{j=1}^{n} \frac{\log f(X_j)}{\log f(X_j \mid \tilde{q}^m)} + o(m). \end{aligned} \tag{5.4.30}$$

Lemma 5.4.8 *Under the conditions of Theorem 5.2.1, there exist two positive constants A and B such that*

$$B m^{\alpha_2} \le \sum_{n_{i,m} > 0} n_{i,m} \log \frac{n_{i,m}}{n \pi_{i,m}} \le A m^{\alpha_1} \quad \text{a.s.} \tag{5.4.31}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$.

Proof. By Taylor expansion,

$$\begin{aligned}
&\sum_{n_{i,m}>0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}} = \sum_{n_{i,m}>0} n_{i,m} \log \left(1 + \frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right) \\
&= \sum_{i=1}^{m} n_{i,m} \left[\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}} - \frac{1}{2}(1+\theta_{i,k})^{-2} \left(\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right)^2\right] \\
&= \sum_{i=1}^{m} \frac{(n_{i,m} - n\pi_{i,m})^2}{n\pi_{i,m}} + \sum_{i=1}^{m} (n_{i,m} - n\pi_{i,m}) \\
&\quad - \sum_{i=1}^{m} \frac{1}{2}(1+\theta_{i,k})^{-2} \left(\frac{(n_{i,m} - n\pi_{i,m})^3}{(n\pi_{i,m})^2} + \frac{(n_{i,m} - n\pi_{i,m})^2}{n\pi_{i,m}}\right) \qquad (5.4.32)
\end{aligned}$$

where $|\theta_{i,k}| \leq \left|\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right|$, so that $\max_{1\leq i\leq m} |\theta_{i,k}| = o(1)$ a.s. uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$. The argument is similar to that used to establish (5.4.24). By Lemma 5.4.4, Lemma 5.4.5, the property $\pi_{i,m} \geq b_1 c_1 m^{-\alpha_1}$ and the following inequality obtained from (5.4.32)

$$\begin{aligned}
&\left|\sum_{n_{i,m}>0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}}\right| \\
&\leq \sum_{i=1}^{m} \frac{(n_{i,m} - n\pi_{i,m})^2}{n\pi_{i,m}} \left[1 + \frac{1}{2}(1+\theta_{i,k})^{-2} \left(1 + \max_{1\leq i\leq m} \left|\frac{n_{i,m} - n\pi_{i,m}}{n\pi_{i,m}}\right|\right)\right], \qquad (5.4.33)
\end{aligned}$$

the lemma can easily be established. □

The following lemma can similarly be proved by Taylor expansion, Lemma 5.4.3, Lemma 5.4.4 and Corollary 5.4.1.

Lemma 5.4.9 *Under the conditions of Theorem 5.2.1,*

$$(1) \qquad \sum_{n_{i,m}>0} \frac{1}{n_{i,m}} = o(m) \quad a.s. \qquad (5.4.34)$$

$$(2) \qquad \sum_{i=1}^{m} \log \frac{n_{i,m}+1}{n\pi_{i,m}+1} = o(m) \quad a.s. \qquad (5.4.35)$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ *as* $n \to \infty$.

Lemma 5.4.10 *Under the conditions of Theorem 5.2.1, there exists a positive constant A such that*

$$-Am^{\alpha_1} + (\alpha_2 - \alpha_1) m \log m + O(m) \leq -L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j \mid \tilde{q}^m)$$

$$\leq (\alpha_1 - 1)m \log m + O(m) \quad a.s. \tag{5.4.36}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ *as* $n \to \infty$.

Proof. First note that

$$-L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j \mid \tilde{q}^m) = \sum_{i=1}^{m} n_{i,m} \log \frac{\pi_{i,m}}{\tilde{r}_{i,m}} - \sum_{i=1}^{m} (n_{i,m}+1) \frac{n_{i,m}+1}{(n+m)\tilde{r}_{i,m}}$$

$$= \sum_{i=1}^{m} \log \tilde{r}_{i,m} + \sum_{i=1}^{m} n_{i,m} \log \frac{(n+m)\pi_{i,m}}{n_{i,m}+1} + \sum_{i=1}^{m} \log \frac{n\pi_{i,m}+1}{n_{i,m}+1}$$

$$+ m \log(n+m) - \sum_{i=1}^{m} \log(n\pi_{i,m}+1). \tag{5.4.37}$$

The second term of the right hand side of (5.4.37)

$$\sum_{i=1}^{m} n_{i,m} \log \frac{(n+m)\pi_{i,m}}{n_{i,m}+1} = \sum_{i=1}^{m} n_{i,m} \log \left[\frac{(n+m)\pi_{i,m}}{n_{i,m}+1} \cdot \frac{n_{i,m}}{n\pi_{i,m}} \cdot \frac{n\pi_{i,m}}{n_{i,m}} \right]$$

$$= n \log \left(1 + \frac{m}{n}\right) - \sum_{n_{i,m}>0} n_{i,m} \log \left(1 + \frac{1}{n_{i,m}}\right) - \sum_{n_{i,m}>0} n_{i,m} \log \frac{n_{i,m}}{n\pi_{i,m}} \tag{5.4.38}$$

and

$$\sum_{n_{i,m}>0} n_{i,m} \log \left(1 + \frac{1}{n_{i,m}}\right) = \sum_{n_{i,m}>0} n_{i,m} \left(\frac{1}{n_{i,m}} + \frac{1}{2}(1+\eta_{i,m})^{-2} \frac{1}{{n_{i,m}}^2} \right)$$

where $0 \leq \eta_{i,m} \leq 1$. By Lemma 5.4.8 and Lemma 5.4.9 (1) we have

$$-Am^{\alpha_1} + O(m) \leq \sum_{i=1}^{m} n_{i,m} \log \frac{(n+m)\pi_{i,m}}{n_{i,m}+1} \leq O(m) \quad \text{a.s.} \tag{5.4.39}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$.

It can also be seen easily that

$$\alpha_2 m \log m + O(m) \leq -\sum_{i=1}^{m} \log \left(\pi_{i,m} + \frac{1}{n}\right) \leq \alpha_1 m \log m + O(m) \tag{5.4.40}$$

From (5.4.25), (5.4.39), Lemma 5.4.9 (2) and (5.4.40) it follows that

$$-Am^{\alpha_1} + (\alpha_2 - \alpha_1) m \log m + O(m) \leq -L_1^*(X^n; m) + \sum_{j=1}^{n} \log f(X_j \mid \tilde{q}^m)$$

$$\leq (\alpha_1 - 1) m \log m + O(m) \quad \text{a.s.} \tag{5.4.41}$$

uniformly in $m \in [n^{\gamma_1}, n^{\gamma_2}]$ as $n \to \infty$. □

Lemma 5.4.11 *Under the conditions (i) — (iv) of Theorem 5.2.2 and* $f \not\equiv 1$, *we have as* $m \to \infty$

$$E_f \log \frac{f}{f(\cdot \mid \tilde{q}^m)} = \sum_{i=1}^{m} \frac{1}{24} \tilde{r}_{i,m}{}^2 \int_{\tilde{Q}_{i,m}} \frac{\dot{f}^2}{f} + o(m^{-2\alpha_2}) \tag{5.4.42}$$

Proof. By the definition of $f(x \mid \tilde{q}^m)$

$$\lim_{m\to\infty} (f(x) - f(x \mid \tilde{q}^m)) = \lim_{m\to\infty} \frac{1}{\tilde{r}_{i,m}(x)} \int_{\tilde{Q}_{i,m}(x)} (f(x) - f(y))dy = 0 \tag{5.4.43}$$

uniformly in $x \in [s,t]$, where $\tilde{Q}_{i,m}(x)$ is the subinterval holding x, and $\tilde{r}_{i,m}(x)$ is the corresponding width. Now by Taylor expansion

$$\begin{aligned} E_f \log \frac{f}{f(\cdot \mid \tilde{q}^m)} &= \sum_{i=1}^{m} \int_{\tilde{Q}_{i,m}} f \log \left(1 + \frac{f - f(\cdot \mid \tilde{q}^m)}{f(\cdot \mid \tilde{q}^m)}\right) \\ &= \sum_{i=1}^{m} \int_{\tilde{Q}_{i,m}} f \cdot \frac{f - f(\cdot \mid \tilde{q}^m)}{f(\cdot \mid \tilde{q}^m)} - \frac{1}{2} \sum_{i=1}^{m} \int_{\tilde{Q}_{i,m}} f(1+\eta_i)^{-2} \left(\frac{f - f(\cdot \mid \tilde{q}^m)}{f(\cdot \mid \tilde{q}^m)}\right)^2 \end{aligned} \tag{5.4.44}$$

where $|\eta_i(x)| \le \left|\frac{f - f(\cdot|\tilde{q}^m)}{f(\cdot|\tilde{q}^m)}\right|$ and by (5.4.43) $\sup_x |\eta_i(x)| = o(1)$. Hence

$$\begin{aligned} E_f \log \frac{f}{f(\cdot \mid \tilde{q}^m)} &= \sum_{i=1}^{m} \int_{\tilde{Q}_{i,m}} \frac{(f - f(\cdot \mid \tilde{q}^m))^2}{f(\cdot \mid \tilde{q}^m)} \\ -\frac{1}{2}(1+o(1)) \sum_{i=1}^{m} \int_{\tilde{Q}_{i,m}} \frac{(f - f(\cdot \mid \tilde{q}^m))^3}{f^2(\cdot \mid \tilde{q}^m)} &- \frac{1}{2}(1+o(1)) \sum_{i=1}^{m} \frac{(f - f(\cdot \mid \tilde{q}^m))^2}{f(\cdot \mid \tilde{q}^m)} \\ &= (1+o(1)) \sum_{i=1}^{m} \frac{1}{2} \int_{\tilde{Q}_{i,m}} \frac{(f - f(\cdot \mid \tilde{q}^m))^2}{f(\cdot \mid \tilde{q}^m)} \end{aligned} \tag{5.4.45}$$

Now by applying the technique used in Proposition 2.7 of Freedman and Diaconis (1981) to prove that

$$\sum_{i=1}^{m} \int_{\tilde{Q}_{i,m}} \frac{(f - f(\cdot \mid \tilde{q}^m))^2}{f(\cdot \mid \tilde{q}^m)} = \frac{1}{12} \sum_{i=1}^{m} \tilde{r}_{i,m}{}^2 \int_{\tilde{Q}_{i,m}} \frac{\dot{f}^2}{f} + o(m^{-2\alpha_2}) \tag{5.4.46}$$

and by (5.4.46) the lemma follows.

By denoting $z = x - \tilde{q}_{i-1,m}$ we have

$$\int_{\tilde{Q}_{i,m}} \frac{(f(x)-f(x \mid \tilde{q}^m))^2}{f(x \mid \tilde{q}^m)}dx = \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}[f(z+\tilde{q}_{i-1,m})-f(z+\tilde{q}_{i-1,m}\mid\tilde{q}^m)]^2dz$$

$$= \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\left[\int_0^z \dot{f}(y+\tilde{q}_{i-1,m})dy - \frac{1}{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-y)\dot{f}(y+\tilde{q}_{i-1,m})dy\right]^2 dz$$

$$= \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\left[\int_0^z \dot{f}(y+\tilde{q}_{i-1,m})dy\right]^2 dz - \frac{1}{\pi_{i,m}}\left[\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-y)\dot{f}(y+\tilde{q}_{i-1,m})dy\right]^2$$

$$= \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^z\int_0^z \dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudvdz$$

$$-\frac{1}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-u)(\tilde{r}_{i,m}-v)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv$$

$$= \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-u\vee v)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv$$

$$-\frac{1}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(\tilde{r}_{i,m}-u)(\tilde{r}_{i,m}-v)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv$$

$$= \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v - \frac{1}{\tilde{r}_{i,m}}uv)\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m})dudv \qquad (5.4.47)$$

where $u \vee v = \max(u,v)$ and $u \wedge v = \min(u,v)$. Direct computation shows that

$$\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v - \frac{1}{\tilde{r}_{i,m}}uv)dudv = \frac{1}{12}\tilde{r}_{i,m}{}^3. \qquad (5.4.48)$$

Define $\overline{\dot{f}}_{i,m} = \frac{1}{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}\dot{f}(u+\tilde{q}_{i-1,m})du$. By (5.4.47)

$$\sum_{i=1}^m \int_{\tilde{Q}_{i,m}} \frac{(f-f(\cdot\mid\tilde{q}^m))^2}{f(\cdot\mid\tilde{q}^m)}$$

$$= \sum_{i=1}^m \frac{\tilde{r}_{i,m}}{\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}\int_0^{\tilde{r}_{i,m}}(u\wedge v - \frac{1}{\tilde{r}_{i,m}}uv)(\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m}) - \overline{\dot{f}}^2_{i,m})dudv$$

$$+\sum_{i=1}^m \frac{\tilde{r}_{i,m}{}^3}{12\pi_{i,m}}\int_0^{\tilde{r}_{i,m}}(\overline{\dot{f}}^2_{i,m} - \dot{f}^2(u+\tilde{q}_{i-1,m}))du$$

$$+\sum_{i=1}^m \frac{\tilde{r}_{i,m}{}^2}{12}\int_0^{\tilde{r}_{i,m}}\left(\frac{\tilde{r}_{i,m}\dot{f}^2(u+\tilde{q}_{i-1,m})}{\pi_{i,m}} - \frac{\dot{f}^2(u+\tilde{q}_{i-1,m})}{f(u+\tilde{q}_{i-1,m})}\right)du$$

$$+\frac{1}{12}\sum_{i=1}^m \tilde{r}_{i,m}{}^2\int_{\tilde{Q}_{i,m}}\frac{\dot{f}^2(x)}{f(x)}dx. \qquad (5.4.49)$$

Note that $|u \wedge v - \frac{1}{\tilde{r}_{i,m}} uv| \leq \tilde{r}_{i,m}$ and

$$\left|\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m}) - \overline{\dot{f}}^2_{i,m}\right| \leq$$
$$\left|\dot{f}(u+\tilde{q}_{i-1,m}) - \overline{\dot{f}}_{i,m}\right| \left|\dot{f}(v+\tilde{q}_{i-1,m})\right| + \left|\dot{f}(v+\tilde{q}_{i-1,m}) - \overline{\dot{f}}_{i,m}\right| \left|\overline{\dot{f}}_{i,m}\right|,$$

then

$$\begin{aligned}
&\left|\sum_{i=1}^{m} \frac{\tilde{r}_{i,m}}{\pi_{i,m}} \int_0^{\tilde{r}_{i,m}} \int_0^{\tilde{r}_{i,m}} (u \wedge v - \frac{1}{\tilde{r}_{i,m}} uv)(\dot{f}(u+\tilde{q}_{i-1,m})\dot{f}(v+\tilde{q}_{i-1,m}) - \overline{\dot{f}}^2_{i,m}) du dv\right| \\
&\quad \leq c_1^{-1} \sum_{i=1}^{m} \tilde{r}_{i,m} \int_0^{\tilde{r}_{i,m}} \left|\dot{f}(u+\tilde{q}_{i-1,m}) - \overline{\dot{f}}_{i,m}\right| \int_0^{\tilde{r}_{i,m}} \left|\dot{f}(v+\tilde{q}_{i-1,m})\right| \\
&\qquad + c_1^{-1} \sum_{i=1}^{m} \tilde{r}_{i,m} \int_0^{\tilde{r}_{i,m}} \left|\dot{f}(v+\tilde{q}_{i-1,m}) - \overline{\dot{f}}_{i,m}\right| \int_0^{\tilde{r}_{i,m}} \left|\overline{\dot{f}}_{i,m}\right| \\
&\quad \leq 2c_1^{-1} \sum_{i=1}^{m} \tilde{r}_{i,m} \int_{\tilde{Q}_{i,m}} \left|\dot{f} - \overline{\dot{f}}_{i,m}\right| \int_{\tilde{Q}_{i,m}} \left|\dot{f}\right|. \qquad (5.4.50)
\end{aligned}$$

Using the Cauchy-Schwartz inequality

$$\begin{aligned}
&\sum_{i=1}^{m} \tilde{r}_{i,m} \int_{\tilde{Q}_{i,m}} \left|\dot{f} - \overline{\dot{f}}_{i,m}\right| \int_{\tilde{Q}_{i,m}} \left|\dot{f}\right| \\
&\leq \left[\sum_{i=1}^{m} \tilde{r}_{i,m} \left(\int_{\tilde{Q}_{i,m}} \left|\dot{f} - \overline{\dot{f}}_{i,m}\right|\right)^2\right]^{\frac{1}{2}} \left[\sum_{i=1}^{m} \tilde{r}_{i,m} \left(\int_{\tilde{Q}_{i,m}} \left|\dot{f}\right|\right)^2\right]^{\frac{1}{2}} \\
&\leq \left[\sum_{i=1}^{m} \tilde{r}_{i,m}{}^2 \int_{\tilde{Q}_{i,m}} \left|\dot{f} - \overline{\dot{f}}_{i,m}\right|^2\right]^{\frac{1}{2}} \left[\sum_{i=1}^{m} \tilde{r}_{i,m}{}^2 \int_{\tilde{Q}_{i,m}} \left|\dot{f}\right|^2\right]^{\frac{1}{2}} \\
&\leq cm^{-2\alpha_2} \left(\int_{[s,t]} (\dot{f} - \overline{\dot{f}}_{i,m})^2\right)^{\frac{1}{2}}
\end{aligned}$$

where $c = (t-s)b_2^2 c_3$ is a constant. By (2.5) of Freedman and Diaconis (1981)

$$\int_{[s,t]} (\dot{f} - \overline{\dot{f}}_{i,m})^2 \to 0 \quad \text{as } m \to \infty. \qquad (5.4.51)$$

Therefore the first term in the right hand side of (5.4.49) is bounded by $o(m^{-2\alpha_2})$. Using the Cauchy-Schwartz inequality, (5.4.51) and a result similar to (5.4.51) for

$$\int_{[s,t]} \left(f - \frac{\pi_{i,m}}{\tilde{r}_{i,m}}\right)^2 \to 0 \quad \text{as } m \to \infty, \qquad (5.4.52)$$

it is easy to show that the second and the third term of the right hand side of (5.4.49) are bounded by $o(m^{-2\alpha_2})$. Therefore (5.4.46) is true, and so is the lemma. □

Lemma 5.4.12 *Under the conditions (i) — (iv) of Theorem 5.2.2, we have*

$$\sum_{j=1}^{n} \log \frac{f(X_j)}{f(X_j \mid \tilde{q}^m)} = nE_f \log \frac{f}{f(\cdot \mid \tilde{q}^m)} + o(nm^{-2\alpha} + m \log n) \quad a.s. \tag{5.4.53}$$

as $n \to \infty$ *uniformly for* $m \in [n^{\gamma_1}, n^{\gamma_2}]$, *where* α *is a constant satisfying* $\alpha_2 \le \alpha < \alpha_2 + \frac{1}{2}$.

Proof. Denote $Z_{j,m} = \log \frac{f(X_j)}{f(X_j|\tilde{q}^m)}$ for each X_j, then $Z_{j,m}$'s are i.i.d. and

$$|Z_{j,m}| \le \max_x \frac{|f - f(\cdot \mid \tilde{q}^m)|}{f(\cdot \mid \tilde{q}^m)} \le \frac{c_3}{c_1} \max_{1 \le i \le m} \tilde{r}_{i,m}.$$

Thus

$$|Z_{j,m} - EZ_{j,m}| \le \frac{2c_3}{c_1} \max_{1 \le i \le m} \tilde{r}_{i,m} \stackrel{\text{def}}{=} B,$$

and

$$\sum_{j=1}^{n} V(Z_{j,m}) \le 4n \frac{c_3^2}{c_1^2} \max_{1 \le i \le m} \tilde{r}_{i,m}{}^2 \stackrel{\text{def}}{=} V.$$

By Bernstein's inequality, for arbitrary $\varepsilon > 0$

$$P\left(\left|\sum_{j=1}^{n}(Z_{j,m} - EZ_{j,m})\right| > \eta\right) \le 2\exp\left\{-\frac{\eta^2}{2(V + \frac{1}{3}B\eta)}\right\}, \tag{5.4.54}$$

where $\eta = n(m^{-2\alpha} + mn^{-1}\log n)\varepsilon$ and $\alpha_2 \le \alpha < \alpha_2 + \frac{1}{2}$. By the definition of B and V,

$$\begin{aligned} V + \frac{1}{3}B\eta &= 4n\frac{c_3^2}{c_1^2} \max_{1 \le i \le m} \tilde{r}_{i,m}{}^2 + \frac{2c_3}{3c_1} \max_{1 \le i \le m} \tilde{r}_{i,m} n(m^{-2\alpha} + mn^{-1}\log n)\varepsilon \\ &\le c'nm^{-2\alpha_2} + c''m^{1-\alpha_2}\log n \end{aligned}$$

where c' and c'' are constants not depending on n and m.

Therefore,

$$\begin{aligned} \frac{\eta^2}{V + \frac{1}{3}B\eta} &\ge \frac{1}{2}\frac{n^2(m^{-2\alpha} + mn^{-1}\log n)^2\varepsilon^2}{\max\{c'nm^{-2\alpha_2}, c''m^{1-\alpha_2}\log n\}} \\ &= \min\{c'n(m^{-2\alpha+\alpha_2} + m^{1+\alpha_2}n^{-1}\log n)^2, \\ &\qquad c''n^2(\log n)^{-1}(m^{-2\alpha-\frac{1}{2}+\frac{1}{2}\alpha_2} + m^{\frac{1}{2}+\frac{1}{2}\alpha_2}n^{-1}\log n)^2\} \end{aligned}$$

for any $m \in [n^{\gamma_1}, n^{\gamma_2}]$ and hence

$$\frac{\eta^2}{V + \frac{1}{3}B\eta} \geq O\left(n^{\frac{-2\alpha+\alpha_2+1}{2\alpha+1}}(\log n)^{\frac{4\alpha-2\alpha_2}{2\alpha+1}}\right). \tag{5.4.55}$$

By (5.4.55) and (5.4.54), it follows that

$$\sum_{n=1}^{\infty} \sum_{m\in[n^{\gamma_1},n^{\gamma_2}]} P\left(\left|\sum_{j=1}^{n}(Z_{j,m} - EZ_{j,m})\right| > \eta\right)$$
$$\leq 2\sum_{n=1}^{\infty} \sum_{m\in[n^{\gamma_1},n^{\gamma_2}]} \exp\left\{-O\left(n^{\frac{-2\alpha+2\alpha_2+1}{2\alpha+1}}(\log n)^{\frac{4\alpha-2\alpha_2}{2\alpha+1}}\right)\right\} < \infty.$$

From the Borel-Cantelli lemma, (5.4.53) follows. □

Proof of Theorem 5.2.2

The first part of the theorem, i.e. the equation (5.2.18) can be obtained from (5.4.30), Lemma 5.4.10, Lemma 5.4.11 and Lemma 5.4.12 and then the second part is straightforward from Theorem 5.2.1. □

Proof of Theorem 5.2.3

$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{(\alpha_1 - 1)m\log m + C_f n m^{-2\alpha_2}\} = M_2 n^{\frac{1}{1+2\alpha_2}}(\log n)^{\frac{2\alpha_2}{1+2\alpha_2}}, \tag{5.4.56}$$
$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\{-Am^{\alpha_1} + (\alpha_2 - \alpha_1)m\log m\} = -M_1(n^{\alpha_1\gamma_2} + n^{\gamma_2}\log n), \tag{5.4.57}$$

the first part is obvious from Theorem 5.2.2. The second part can be established along similar lines. □

Proof of Theorem 5.2.4.

Regarding m as a real value and taking the derivative of $\frac{1}{2}n\log\frac{n}{m} + C'_f n m^{-2}$ with respect to m, we get

$$\min_{m\in[n^{\gamma_1},n^{\gamma_2}]}\left\{\frac{1}{2}n\log\frac{n}{m} + C'_f n m^{-2}\right\} = M_5 n^{\frac{1}{3}}(\log n)^{\frac{2}{3}} \tag{5.4.58}$$

and the minimization is achieved at $m = M_6(n/\log n)^{\frac{1}{3}}$. By this result and Theorem 5.2.2, (a), (b), (c) and (d) are readily obtained. □

Proof of Theorem 5.3.1.

As in Lemma 5.4.7, it can be shown that

$$L_4(\tilde{q}_1^{m_1}, \cdots, \tilde{q}_k^{m_k}, m_1, \cdots, m_k, \delta) = o\left(\sum_{i=1}^{k} m_i\right). \tag{5.4.59}$$

If either $\alpha_1 \neq 1$ or $\alpha_2 \neq 1$, then by Theorem 5.2.3

$$-M_3 \sum_{i=1}^{k}(n_i^{\alpha_1\gamma_2} + n_i^{\gamma_2}\log n_i) \leq C(X_1^{n_1}, \cdots, X_k^{n_k}) + \sum_{i=1}^{k} \log f_i^{n_i}(X_i^{n_i})$$

$$\leq M_4 \sum_{i=1}^{k} n_i^{\frac{1}{1+2\alpha_2}} (\log n_i)^{\frac{2\alpha_2}{1+2\alpha_2}} \quad \text{a.s.} \tag{5.4.60}$$

and

$$-M_3(n^{\alpha_1\gamma_2} + n^{\gamma_2}\log n) \leq C(X^n) + \log f_{mix}^n(X^n) \leq M_4 n^{\frac{1}{1+2\alpha_2}} (\log n)^{\frac{2\alpha_2}{1+2\alpha_2}} \quad \text{a.s.} \tag{5.4.61}$$

for some positive constants M_3 and M_4 depending on $f_1, \cdots, f_k$.

If $\alpha_1 = \alpha_2 = 1$, then by Theorem 5.2.4 (b)

$$C(X_1^{n_1}, \cdots, X_k^{n_k}) + \sum_{i=1}^{k} \log f_i^{n_i}(X_i^{n_i}) = O\left(\sum_{i=1}^{k} n_i^{\frac{1}{3}} (\log n_i)^{\frac{2}{3}}\right) \quad \text{a.s.} \tag{5.4.62}$$

and

$$C(X^n) + \log f_{mix}^n(X^n) = O\left(n^{\frac{1}{3}}(\log n)^{\frac{2}{3}}\right) \quad \text{a.s..} \tag{5.4.63}$$

It remains to prove that there exists a constant $\eta < 0$ such that

$$\frac{1}{n}\left(\log f_{mix}^n(X^n) - \sum_{i=1}^{k} \log f_i^{n_i}(X_i^{n_i})\right) < \eta \quad \text{a.s.} \tag{5.4.64}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$ satisfying $\frac{n_1}{n} > \varepsilon_1 > 0, \cdots, \frac{n_k}{n} > \varepsilon_k > 0$ for any prescribed constants $\varepsilon_1, \cdots, \varepsilon_k$, if at least two of $f_1, \cdots, f_k$ are not equal almost surely, and

$$\frac{1}{n}\left(\log f_{mix}^n(X^n) - \sum_{i=1}^{k} \log f_i^{n_i}(X_i^{n_i})\right) \to 0 \quad \text{a.s.} \tag{5.4.65}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$ if $f_1 = f_2 = \cdots = f_k$ a.s..

Because

$$\sum_{i=1}^{k} \log f_i^{n_i}(X_i^{n_i}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \log f_i(X_{ij}),$$

$$\log f_{mix}^n(X^n) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \log\left((\sum_{l=1}^{k} \frac{n_l}{n} f_l(X_{ij})\right)$$

and f_i's are bounded density functions, by the strong law of large numbers for i.i.d. random variables it follows that

$$\frac{1}{n}\sum_{i=1}^{k}\log f_i^{n_i}(X_i^{n_i}) - \sum_{i=1}^{k}\frac{n_i}{n}\int f_i \log f_i \to 0 \quad \text{a.s.} \tag{5.4.66}$$

and

$$\frac{1}{n}\log f_{mix}^{n}(X^n) - \int f_{mix}\log f_{mix} \to 0 \quad \text{a.s.} \tag{5.4.67}$$

as $n_1 \to \infty, \cdots, n_k \to \infty$. By the convexity of $x\log x$,

$$\int f_{mix}\log f_{mix} \le \sum_{i=1}^{k}\frac{n_i}{n}\int f_i \log f_i \tag{5.4.68}$$

for any group of samples of sizes $n_1, \cdots, n_k$ satisfying $\sum_{i=1}^{k} n_i = n$, where the equality holds if and only if all the densities $f_1, \cdots, f_k$ are equal (except a set of measure zero). Therefore (5.4.65) is established by using (5.4.66) and (5.4.67). Also for any $\varepsilon_1 > 0, \cdots \varepsilon_k > 0$ if $\frac{n_1}{n} > \varepsilon_1, \cdots, \frac{n_k}{n} > \varepsilon_k$, and if at least two of $f_1, \cdots, f_k$ are not equal a.s. almost surely, there exists a constant $\eta < 0$ depending on $\varepsilon_1, \cdots, \varepsilon_k$ such that

$$\int f_{mix}\log f_{mix} - \sum_{i=1}^{k}\frac{n_i}{n}\int f_i \log f_i < \eta \tag{5.4.69}$$

for any set of integers $\{n_i\}$ satisfying $\sum_{i=1}^{k} n_i = n$. Hence (5.4.64) follows from (5.4.66) and (5.4.67). Notice that $\alpha_1\gamma_2 < 1$, $\gamma_2 < 1$ and $\frac{1}{1+2\alpha_2} < 1$, (5.3.17) and (5.3.18) hold by (5.4.59) to (5.4.65). □

Chapter 6

Concluding Remarks

6.1 Summary

In Chapter 2 we proposed an index of predictive power as a criterion to select the principal components of a random vector distributed in a parametric family. This criterion, when applied to the principal components selection, considers the lost information due to the reduction of the parameters as well as the observed variables. The principal components, obtained by minimizing the index of predictive power, turn out to be identical to the classical principal components when the assumed distribution is normal. A test procedure for the principal components selection was constructed and discussed. Finally, principal components for a type of ε-contaminated normal family were given.

In Chapter 3 we considered the problem of selecting a model with the best predictive ability in a class of generalized linear models. A predictive least quasi-deviance criterion was proposed to measure the predictive ability of a model. This criterion is obtained by applying the idea of the predictive minimum description length principle and the theory of quasi-likelihood functions. The resulting predictive quasi-deviance function is an extension of the predictive stochastic complexity of the model. Under rather weak conditions the predictive least quasi-deviance method was shown to be consistent in the sense that the probability of selecting the right model converges to one as the number of observations goes to infinity. Also we have shown that the

selected model converges to the optimal model in expectation. The method was then modified for finite sample applications. Justifications and discussions were provided and examples and simulation results were presented.

In Chapter 4 a density estimation based complexity decision rule was proposed which uses the quality of these estimators to estimate the corresponding unknown element of the true probability density. In the development we introduced a loss function which includes the total variation of the squared distance of the characteristic functions to evaluate the performance of the density decision rule. The resulting complexity density decision procedure was shown to be admissible, to achieve the minimum expected risk, and to form a minimal complete class.

In Chapter 5, a generalized histogram density estimator with unequal-width subintervals was used to find both optimal and predictive optimal description of a data sample. Both optimal descriptions were expressed in terms of Rissanen's stochastic complexity. Uniform almost sure asymptotic expressions for both descriptions were given. Finally, as an application of a stochastic complexity for optimal data description, a new test procedure for hypothesis of homogeneity was proposed and proved to have an asymptotic power 1 in the limit. Examples and simulation results are also supplied.

6.2 Future Research

There still remains a great deal of work to develop the stochastic complexity as a competent method in statistics inference.

In ordinary linear regression a model selection criterion by stochastic complexity is called the predictive least square principle (PLS). In the case of i.i.d. normal residuals the PLS principle is known to be consistent. It is important to study the effects of small deviation from independence to the PLS principle. For instance, when the regression residuals come from a Gaussian stationary process with the long range dependence structure, it is interesting to know whether PLS is still consistent and whether it is still as efficient as in the i.i.d. case. Only when the behavior of this

simple regression case is clear, it becomes possible to study the effect of long range or other types of dependence on more complex modeling problems.

The study of principal components selection from parametric point of view may be extended to a nonparametric standpoint. For example we can define an empirical distribution, calculate the stochastic complexity of a vector variable with large dimension based on that distribution, then formulate the index of predictive power and conduct the principal components selection based on this index.

The fundamental idea in Chapter 3 is using the accumulated prediction error as a model selection criterion. This may be applied naturally to other regression problems, such as the regression using splines and polynomials, nonparametric regression and additive regression, etc.

As it was noted in Chapter 4, it is possible to find an application of stochastic complexity theory in finite decision-problems (identification). It is also possible to derive a nonparametric density estimation based complexity decision rule and study the properties of admissibility and completeness for this decision rule.

In Chapter 5 we have shown the power of using stochastic complexity to find an optimal histogram density estimation and to proceed with other selection problems associated with the histogram density. This contrasts with the usual way of assessing density estimates, either subjectively or by their asymptotic properties. Knowing that the stochastic complexity provides a global measure for evaluating the success of modeling reality through an observed data string, we may tackle other nonparametric density and curve estimation problems and their possible applications.

Appendix A

Programs for Chapter 3

```
c     This is the program for Example 3.5.1. This program is used
c     to select the optimal model and compute the probability of
c     selecting the true model by using monte carlo PLQD method.
c     It is valid for linear regression problems.

      implicit double precision (a-h,o-z)
      parameter(maxr=1000,maxc=20,maxt=100)
      dimension x(40,5), xp(40,5)
      dimension y(40,1000),yp(40,1000),dum(16),mint(1000)
      dimension index(40),salpha(1000,16),sdev(1000,16)
      dimension model(16,5),modtr(5),coeff(40,5,1000)
      character*50 infile1, infile2
      common model,coeff

      write(*,*) 'Input the true model'
      read(*,*) (modtr(i),i=1,5)
      write(*,*) 'Input the data file of independent variables'
      read(*,*) infile1
      open(15,file=infile1,status='old')
```

```
 10   write(*,*) 'Input the number of independent variables'
      read(*,*) nx
      if(nx.gt.maxc) goto 10
      write(*,*) 'Input the number of data points'
      read(*,*) ndata
      write(*,*) 'Input the data file of response values'
      read(*,*) infile2
      open(19,file=infile2,status='old')
      write(*,*) 'Input the number of data points used to do the
     &    first regression'
      read(*,*) nd
      write(*,*) 'Input the number of monte carlo simulations'
      read(*,*) nb
      write(*,*) 'Input the number of response variables'
      read(*,*) ny

      do ii=1,ndata
        read(15,*) (x(ii,j),j=1,nx+1)
        read(19,*) (y(ii,k),k=1,ny)
        index(ii)=ii
      enddo
c     set up a non-repeatable initial state for permutation
      call g05ccf
c     'salpha' contains the PLQD values for each response
c     data in each model
      do i=1,ny
        do j=1,2**nx
          salpha(i,j)=0.d0
        enddo
```

```
      enddo

      do 20 ib=1,nb
c     set up a permutation of index
      call g05ehf(index,ndata,ifail)
c     write(*,*) 'index is'
c     write(*,*) index
c     creat the corresponding permutation of independent variable
c     data and response data
      do i=1,ndata
        do j=1,nx+1
          xp(i,j)=x(index(i),j)
        enddo
        do k=1,ny
          yp(i,k)=y(index(i),k)
        enddo
      enddo
      call predev(xp,yp,nd,ndata,nx,ny,sdev)
      do i=1,ny
        do j=1, 2**nx
          salpha(i,j)=salpha(i,j)+sdev(i,j)/real(nb)
        enddo
      enddo
 20   continue
c     write(*,*) 'salpha is'
c     do i=1,ny
c     write(*,*) (salpha(i,j),j=1,2**nx)
c     enddo
      ksum=0
```

```
      write(*,*) 'the best model is'
      do iy=1,ny
        do j=1,2**nx
          dum(j)=salpha(iy,j)
        enddo
        mint(iy)=indexmin(dum,2**nx)
        ks=0
        do k=1,nx+1
          if(modtr(k).ne.model(mint(iy),k)) then
          ks=ks+1
          end if
        enddo
        if(ks.eq.0) ksum=ksum+1
        write(*,*) (model(mint(iy),j),j=1,nx+1)
      enddo
      write(*,*) 'the probability*1000 is'
c     the probability is the empircal probability of
c     selecting the optimal model.
      write(*,*) ksum
      write(*,*) 'the monte carlo PLQD value are'
      write(*,*) (salpha(i,mint(i)),i=1,ny)
      stop
      end


c     This function is used to find the index where the component
c     of x is minimum.
      function indexmin(x,n)
      implicit double precision(a-h,o-z)
```

```
      dimension x(n)
      temp=9999999999.d0
      ind=-1
      do ii=1,n
        if(x(ii).le.temp) then
        temp=x(ii)
        ind=ii
        endif
      enddo
      indexmin=ind
      return
      end


c     This one selects all the subsets of set {1,2,...n}
      subroutine possmod(n,nsubset)
      logical modmat(1024,10),bit
      common /subs/ modmat
      do i=0,nsubset-1
        do j=n-1,0,-1
          modmat(i+1,n-j)=bit(j,i)
        enddo
      enddo
      return
      end


c     This subroutine is used to find PLQD(sdev) value for (x,y),
c     where x is the matrix contains the x-variables values, and
```

```
c     y is the matrix contains response values.

      subroutine  predev(x,y,nd,ndata,nx,ny,sdev)
      implicit double precision (a-h,o-z)
      parameter(maxr=1000,maxc=20,maxt=100)
      dimension x(40,5),y(40,1000),sdev(1000,16)
      dimension xr(40,5),yr(40,1000)
      dimension model(16,5),coeff(40,5,1000)
      common model,coeff


      do i=1,ny
        do j=1,2**nx
          sdev(i,j)=0.d0
        enddo
      enddo
      do 30 iv=nd,ndata-1
        do ii=1,iv
          do j=1,nx+1
            xr(ii,j)=x(ii,j)
          enddo
          do j=1,ny
            yr(ii,j)=y(ii,j)
          enddo
        enddo
        call coefmod(xr,yr,iv,nx,ny)
        do im=1,2**nx
          do iy=1,ny
           tempv=0.d0
```

```
          do ix=1,nx+1
            tempv=tempv+x(iv+1,ix)*coeff(im,ix,iy)
          enddo
          sdev(iy,im)=sdev(iy,im)+
     &(y(iv+1,iy)-tempv)*(y(iv+1,iy)-tempv)/(2.0*real(ndata-nd))
        enddo
      enddo
 30   continue
      return
      end


c     This subroutine finds the coefficient matrix and
c     all the possible models.

      subroutine coefmod(xr,yr,ndata,nx,ny)
      implicit double precision (a-h,o-z)
      parameter(maxr=1000,maxc=20,maxt=100)
      dimension xr(40,5),yr(40,1000),xt(40,5)
      dimension model(16,5),coeff(40,5,1000)
      dimension sigsq(maxr),C(maxt,maxc),coef(maxc,maxr)
      dimension ipiv(maxc),wk1(maxc,4),wk2(maxt)
      logical modmat(1024,10)
      common /subs/ modmat
      common model,coeff

      call possmod(nx,2**nx)
      do ii=1,2**nx
        do jj=1,nx+1
```

```
        model(ii,jj)=0
      enddo
    enddo
    do 40 im=1,2**nx
      nvar=1
      do kk=1,ndata
        xt(kk,nvar)=xr(kk,nvar)
      enddo
      model(im,1)=1
      do jj=1,nx
        if(modmat(im,jj)) then
          nvar=nvar+1
          model(im,nvar)=jj+1
          do kk=1,ndata
            xt(kk,nvar)=xr(kk,jj+1)
          enddo
        endif
      enddo
      ifail=0
      call g02cjf(xt,40,yr,40,ndata,nvar,maxr,coef,maxc,
     &   sigsq,C,maxt,ipiv,wk1,wk2,ifail)
      do i=1,ny
        do k=1,nx+1
          coeff(im,k,i)=0.d0
        enddo
      enddo
      do iy=1,ny
        do jj=1,nvar
          coeff(im,model(im,jj),iy)=coef(jj,iy)
```

```
          enddo
        enddo
 40   continue
      return
      end
```

==========

```
function(x, y, n)
{
#: This is S-plus program, used in Example 3.5.1
#: It is to find the probability of selecting the optimal model
#: by using approximate PLQD method based on "n" simulations.
#: "x" is the matrix contains columns of explanatory variable values.
#: "y" is the matrix contains "n" columns of response values,
#: generated by standard normal distribution.
len <- nrow(y)
p <- ncol(x) - 1
beta <- c(2, 9, 0, 4, 8)
mdl <- c(1, 0, 3, 4)
y <- y + x %*% t(beta) %*% c(1:n)
s.alpha <- matrix(0, n, 2^p)
for(i in (2 * p + 2):len) {
s.alpha[, 1] <- s.alpha[, 1] + ((y[i, ] -
  apply(y[1:(i - 1),], 2, mean))^2)/(2 * (len - 2 * p - 1))
}
mmodel <- fantas(p)
for(m in 1:(2^p - 1)) {
dum <- sum(mmodel[m, ])
```

```
mrow <- mmodel[m, ]
xx <- x[, 2:(p + 1)][, mrow != 0]
for(k in (2 * p + 1):(len - 1)) {
if(dum == 1) {
coeff <- lsfit(xx[1:k], y[1:k, ])$coef
s.alpha[, (m + 1)] <- s.alpha[, (m + 1)] + (y[(k + 1), ] -
as.numeric(c(1, xx[k + 1]) %*% coeff))^2/(2 * (len - 2 * p - 1))
}
else {
coeff <- lsfit(xx[1:k, ], y[1:k, ])$coef
s.alpha[, (m + 1)] <- s.alpha[, (m + 1)] + (y[(k + 1), ] -
as.numeric(c(1, xx[(k + 1), ])
%*% coeff))^2/(2 * (len - 2 * p - 1))
}
}
}
model.mat <- rbind(c(rep(0, p)), mmodel)
so <- t(apply(s.alpha, 1, sort))
lmod <- matrix(-1, n, 4)
pmod <- lmod
prob <- 0
for(i in 1:n) {
sm <- model.mat[c(1:2^p)[s.alpha[i, ] == so[i, 1]], ]
if(is.vector(sm))
pmod[i, ] <- sm
else pmod[i, ] <- sm[order(apply(sm, 1, sum))[1], ]
if(sum(pmod[i, ]) == 0)
lmod[i, ] <- c(0, 0, 0, 0)
else {
```

```
lmod[i, ][pmod[i, ] != 0] <- c(1:p)[pmod[i, ] != 0]
lmod[i, ][pmod[i, ] == 0] <- c(rep(0, p))[pmod[i, ] == 0]
}
prob <- prob + 1 - abs(sign(sum(lmod[i, ] - md1)))
}
prob <- prob/n
print("the probability of selecting the optimal model is")
print(prob)
print("the best model is ")
print(lmod)
print("the PLQD value is")
so[, 1]
}


#: Splus program "fantas", used to find all
#: the subsets of {1,2,...,p}.
function(p)
{
a <- array(data = 0, c(2^p - 1, p))
if(p <= 1)
a <- 1
else {
a[1, 1] <- 1
a[2:2^(p - 1), 1] <- c(rep(1, 2^(p - 1) - 1))
a[2:2^(p - 1), 2:p] <- fantas(p - 1)
a[(2^(p - 1) + 1):(2^p - 1), 1] <- c(rep(0, 2^(p - 1) - 1))
a[(2^(p - 1) + 1):(2^p - 1), 2:p] <- fantas(p - 1)
}
```

```
a
}
```

```
c     This program is used to select the optimal model and compute
c     the probability of selecting the true model by using monte
c     carlo PLQD method.It is valid for generalized linear regression
c     model with Poisson error. It is used in Example 3.5.2.


c     main program

      implicit double precision (a-h,o-z)
      parameter(maxr=36,maxc=4,maxs=8,maxyc=1000)
      dimension x(maxr,maxc), xp(maxr,maxc)
      dimension y(maxr,maxyc),yp(maxr),dum(maxs),mint(maxyc)
      dimension index(maxr),salpha(maxyc,maxs),sdev(maxs)
      dimension model(maxs,maxc),modtr(maxc)
      character*50 infile1, infile2
      common model

      write(*,*) 'Input the true model'
      read(*,*) (modtr(i),i=1,maxc)
      write(*,*) 'Input the data file of independent variables'
      read(*,*) infile1
c     the values of the first column of infile1 are 1.
      open(15,file=infile1,status='old')
 10   write(*,*) 'Input the number of independent variables'
```

```
      read(*,*) nx
c     'nx' does not count the intercept term in the model.
      if(nx.gt.maxc) goto 10
      write(*,*) 'Input the number of data points'
      read(*,*) ndata
      write(*,*) 'Input the data file of response values'
      read(*,*) infile2
      open(19,file=infile2,status='old')
      write(*,*) 'Input the number of data points used to do the
     & first regression'
      read(*,*) nd
      write(*,*) 'Input the number of monte carlo simulations'
      read(*,*) nb
      write(*,*) 'Input the number of response variables'
      read(*,*) ny

      do ii=1,ndata
        read(15,*) (x(ii,j),j=1,nx+1)
        read(19,*) (y(ii,k),k=1,ny)
        index(ii)=ii
      enddo
c     set up a non-repeatable initial state for permutation
      call g05ccf
c     'salpha' contains the PLQD values for each respcnse
c     data in each model

      ksum=0
      write(*,*) 'the best model is'
      do 105 iy=1,ny
```

```
      do j=1, 2**nx
        salpha(iy,j)=0.d0
      enddo
      do 20 ib=1,nb
c     set up a permutation of index
      call g05ehf(index,ndata,ifail)
c     creat the corresponding permutation of independent variables
c     data and response data
      do i=1,ndata
        do j=1,nx+1
          xp(i,j)=x(index(i),j)
        enddo
        yp(i)=y(index(i),iy)
      enddo
      call predev(xp,yp,nd,ndata,nx,sdev)
      do j=1, 2**nx
        salpha(iy,j)=salpha(iy,j)+sdev(j)/dble(nb)
      enddo
 20   continue

      do j=1,2**nx
        dum(j)=salpha(iy,j)
      enddo
      mint(iy)=indexmin(dum,2**nx)
      ks=0
      do k=1,nx+1
        if(modtr(k).ne.model(mint(iy),k)) then
          ks=ks+1
        endif
```

```
      enddo
      if(ks.eq.0) ksum=ksum+1
      write(*,*) (model(mint(iy),j),j=1,nx+1)
 105  continue

      write(*,*) 'the probability*1000 is'
c     This probability is the empircal probability of
c     selecting the optimal model.
      write(*,*) ksum
      write(*,*) 'the monte carlo PLQD values are'
      write(*,*) (salpha(iy,mint(iy)),iy=1,ny)
      stop
      end



c     This function is used to find the index where the component
c     of x is minimum.
      function indexmin(x,n)
      implicit double precision(a-h,o-z)
      dimension x(n)
      temp=9999999999.d0
      ind=-1
      do ii=1,n
        if(x(ii).le.temp) then
          temp=x(ii)
          ind=ii
        endif
      enddo
```

```
      indexmin=ind
      return
      end



c     This one selects all the subsets of set {1,2,...n}
      subroutine possmod(n,nsubset)
      logical modmat(1024,10),bit
      common /subs/ modmat
      do i=0,nsubset-1
        do j=n-1,0,-1
          modmat(i+1,n-j)=bit(j,i)
        enddo
      enddo
      return
      end



c     This subroutine is used to find PLQD(sdev) values for all
c     the possible models (x,y), where x is the matrix contains
c     the x-variables values, and y is the matrix contains
c     response values.

      subroutine  predev(x,y,nd,ndata,nx,sdev)
      implicit double precision (a-h,o-z)
      parameter(maxr=36,maxc=4,maxs=8)
      dimension x(maxr,maxc),y(maxr),sdev(maxs)
```

```
      dimension xr(maxr,maxc),yr(maxr)
      dimension xt(maxr,maxc),x0(maxr,maxc)
      dimension model(maxs,4),coeff(maxs,4)
      common model


      do j=1,2**nx
        sdev(j)=0.d0
      enddo
      do 30 iv=nd,ndata-1
        do ii=1,iv
          do j=1,nx+1
            xr(ii,j)=x(ii,j)
            xt(ii,j)=x(ii,(j+1))
          enddo
          yr(ii)=y(ii)
        enddo
        call coefmod(xt,yr,iv,nx,coeff,x0)
        do 35 im=1,2**nx
          tempv=0.d0
          do ix=1,nx+1
            tempv=tempv+x(iv+1,ix)*coeff(im,ix)
          enddo
          sdev(im)=sdev(im)+(y(iv+1)*(dlog(y(iv+1))-
     & tempv-1.d0)+dexp(tempv))/dble(ndata-nd)
35    continue
30    continue
      return
      end
```

```
c     This subroutine finds the coefficient matrix and all
c     the possible models.
      subroutine coefmod(xt,yr,ndata,nx,coeff,x0)
      implicit double precision (a-h,o-z)
      parameter(maxr=36,maxc=4,maxs=8)
      dimension yr(ndata),xt(maxr,maxc),x0(ndata,nx)
      dimension model(8,4),coeff(maxs,4)
      dimension isx(maxc),b(maxc),se(maxc)
      dimension cov((maxc+1)*(maxc+2)/2),v(maxr,maxc+8)
      dimension wk(((maxc+1)**2+3*(maxc+1)+22)/2)
      logical modmat(1024,10)
      character link, mean, offset, weight
      common /subs/ modmat
      common model

      call possmod(nx,2**nx)
      do ii=1,2**nx
        do jj=1,nx+1
          model(ii,jj)=0
        enddo
      enddo
      do i=1,ndata
        do j=1,nx
          x0(i,j)=xt(i,j)
        enddo
      enddo
```

```
do 40 im=1,2**nx
  do i=1,maxc
    b(i)=0.d0
  enddo
  do i=1,nx
    isx(i)=0
  enddo
  nvar=0
  model(im,1)=1
  do jj=1,nx
    if(modmat(im,jj)) then
      nvar=nvar+1
      model(im,nvar+1)=jj+1
      isx(jj)=1
    endif
  enddo
  do k=1,nx+1
    coeff(im,k)=0.d0
  enddo
  ifail=-1
  link='l'
  mean='m'
  offset='n'
  weight='u'
  ldx=ndata
  ip=nvar+1
  ldv=ndata
  tol=0.00005d0
  maxit=0
```

```
      iprint=0
      eps=0.000001d0
      call g02gcf(link, mean, offset, weight, ndata,
     &   x0, ldx, nx, isx, ip, yr, wt, a,
     &   dev, idf, b, irank, se, cov, v, ldv, tol,
     &   maxit, iprint, eps, wk, ifail)
      do jj=1,nvar+1
        coeff(im,model(im,jj))=b(jj)
      enddo
 40   continue
      return
      end
```

==========

```
function(x, y, mmodel, bm)
{
# This is the S-plus program for Example 3.5.3 of Chapter 3.
#  x: data matrix contains  observations of explanatory variables.
# y: observations for response variable;
# mmodel: all possible candidate models;
# bm: number of permutations( monte carlo )
len <- length(y)
num <- nrow(mmodel)
s.alpha <- c(rep(0, num))
for(j in 1:bm) {
sam <- sample(len)
y1 <- y[sam]
x1 <- x[sam,  ]
```

```
for(i in 20:23) {
coeff <- as.numeric(glm(y1[1:(i - 1)] ~ 1, binomial,
  maxit = 15, bf.maxit = 15, trace = F)$coef)
pihat <- exp(coeff)/(1 + exp(coeff))
if(y1[i] == 0) {
s.alpha[1] <- s.alpha[1] - (log(1 - pihat))/(4 * bm)
}
else {
if(y1[i] != 1) {
s.alpha[1] <- s.alpha[1] + (y1[i] * log(y1[i]/pihat) + (1 - y1[i])
  * log((1 - y1[i])/(1 - pihat)))/(4 * bm)
}
else {
s.alpha[1] <- s.alpha[1] - (log(pihat))/(4 * bm)
}
}
}
for(m in 2:num) {
mrow <- mmodel[m, ]
xx <- as.matrix(x1[, mrow != 0])
for(k in 20:23) {
coeff <- as.numeric(glm(y1[1:(k - 1)] ~ xx[1:(k - 1), ],
  binomial, maxit = 15, bf.maxit = 15, trace = F)$coef)
coeff.ok <- !is.na(coeff)
muhat <- coeff[coeff.ok] %*% c(1, xx[k, ])[coeff.ok]
pihat <- exp(muhat)/(1 + exp(muhat))
if(y1[k] == 0) {
s.alpha[m] <- s.alpha[m] - (log(1 - pihat))/(4 * bm)
}
```

```
else {
if(y1[k] != 1) {
s.alpha[m] <- s.alpha[m] + ( y1[k] * log(y1[k]/ pihat) +
(1 - y1[k]) * log((1 - y1[k])/(1 - pihat)))/(4 * bm)
}
else {
s.alpha[m] <- s.alpha[m] - (log(pihat))/(4 * bm)
}
}
}
}
}
so <- sort(s.alpha)
otm <- mmodel[s.alpha == so[1], ]
# print("the optimal model is")
# print(otm)
# print("the PLQD value is ")
# print(so[1])
# print("s.alpha is")
s.alpha
}
```

Appendix B

Programs for Chapter 5

```
c     This is a Fortran program for Subsection 5.3.4:
c     Simulation Studies
      implicit double precision (a-h,o-z)
      parameter(maxr=1000,npool=30,k=2)
      integer nrh0,nk(k),maxm(k),maxmp,maxn
      dimension obs(maxr,npool),srp(maxr,2),sr(maxr,2,k)
      dimension opstc(maxr,2),opstk(maxr,k+1)
      dimension obs1(npool), obsn(k,1000),srp1(2),sr1(k,2)
      dimension opstc1(2),opstk1(k+1)
      character*50 dfile1, dfile2, dfile3, dfile4

      write(*,*) 'input the sample sizes'
      read(*,*) (nk(i),i=1,2)
      write(*,*) 'input the maximum numbers of equal-width bins'
      read(*,*) (maxm(i),i=1,2)
      write(*,*) 'input the maximum number of equal-width bins
     & for the pooled sample'
      read(*,*) maxmp
      write(*,*) 'input the digit'
      read(*,*) dig
```

```
 write(*,*) 'input the data file of the observations'
 read(*,*) dfile1
 open(unit=75, file=dfile1,status='old')
 write(*,*) 'input the smallest value and the range
&   of the pooled sample'
 read(*,*) dfile2
 open(unit=79,file=dfile2,status='old')
 write(*,*) 'input the smallest value and the range
&  of the first sample'
 read(*,*) dfile3
 open(unit=81,file=dfile3,status='old')
 write(*,*) 'input the smallest value and the range
&  of the second sample'
 read(*,*) dfile4
 open(unit=83,file=dfile4,status='old')

 do ii=1,maxr
   read(75,*) (obs(ii,j),j=1,npool)
   read(79,*) (srp(ii,jj),jj=1,2)
   read(81,*) (sr(ii,jj,1),jj=1,2)
   read(83,*) (sr(ii,jj,2),jj=1,2)
 enddo

 maxn=max0(nk(1),nk(2))
 nrh0=0
 do i=1,maxr
  do j=1,npool
   obs1(j)=obs(i,j)
  enddo
```

```
      srp1(1)=srp(i,1)
      srp1(2)=srp(i,2)
      do j=1,nk(1)
       obsn(1,j)=obs(i,j)
      enddo
      sr1(1,1)=sr(i,1,1)
      sr1(1,2)=sr(i,2,1)
      do j=1,nk(2)
       obsn(2,j)=obs(i,(nk(1)+j))
      enddo
      sr1(2,1)=sr(i,1,2)
      sr1(2,2)=sr(i,2,2)
      call opms(obs1,srp1,npool,maxmp,dig,opstc1)
      opstc(i,1)=opstc1(1)
      opstc(i,2)=opstc1(2)
      call opms2(obsn,sr1,nk,maxn,maxm,dig,opstk1)
      do jj=1,(k+1)
       opstk(i,jj)=opstk1(jj)
      enddo
      if (opstk(i,1).le.opstc(i,1)) then
       nrh0=nrh0+1
      endif
     enddo

     write(*,*) 'digit=',dig
     write(*,*) 'the number of cases when H0 are rejected'
     write(*,*) nrh0
c    write(*,*) 'ideal codelength under H0, optimal m; codelength
c    & under H1, optimal m'
```

```
c       do i=1,maxr
c        write(*,*) (opstc(i,j),j=1,2),(opstk(i,j),j=1,(k+1))
c       enddo
        stop
        end


        subroutine opms(obs,sr,n,maxm,dig,opstc)
c       Compute the idealized codelength('opstc(1)')
c       (stochastic complexity + minimum description length)
c       for one sample of data.
        implicit double precision (a-h,o-z)
        integer n,maxm
        dimension obs(n),sr(2),opstc(2)

        opstc(1)=-1.d05
        do m=1,maxm
         call stcmpk(obs,sr,m,n,1,n,dig,stc)
         if ((opstc(1).eq.(-1 d05)).or.(opstc(1).gt.stc)) then
          opstc(1)=stc
          opstc(2)=m
         endif
        enddo
        return
        end


        subroutine opms2(obsn,srk,nk,maxn,maxm,dig,opstk)
c       Compute the idealized codelength('opstk(1)') (stochastic
```

```
c     complexity + minimum description length)
c     for two samples of data.
      implicit double precision (a-h,o-z)
      integer nk(2),maxn,maxm(2)
      dimension obsn(2,maxn),srk(2,2),opstk(3),nvec(10000)
      dimension stcc(10000,2),sr(2),para(6)
      external d2lg

      do jk=1,2
       sr(1)=srk(jk,1)
       sr(2)=srk(jk,2)
       do m=1,maxm(jk)
        stcc(m,jk)=0.d0
        nvec(1)=nk(jk)
        if (m.gt.1) then
         nvec(1)=0
         do i=1,nk(jk)
          if ((obsn(jk,i).ge.sr(1)).and.(obsn(jk,i).l..
     &       (sr(1)+(1.d0/m)*sr(2)))) then
           nvec(1)=nvec(1)+1
          endif
         enddo
        nvec(m)=0
         do i=1,nk(jk)
          if (obsn(jk,i).gt.(sr(1)+((m-1)/dble(m))
     &        *sr(2))) then
           nvec(m)=nvec(m)+1
          endif
         enddo
```

```
          if (m.gt.2) then
           do j=2,(m-1)
            nvec(j)=0
            do i=1,nk(jk)
             if ((obsn(jk,i).gt.(sr(1)+((j-1)/dble(m))
     &          *sr(2))).and.(obsn(jk,i).le.(sr(1)
     &          +(j/dble(m))*sr(2)))) then
              nvec(j)=nvec(j)+1
             endif
            enddo
           enddo
          endif
         endif
c         write(*,*) 'nvec', (nvec(j),j=1,m)
         call cplxty(sr,nk(jk),m,nvec,cplx)
         stcc(m,jk)=cplx
        enddo
        para(3*jk-2)=dsign(dint(dabs(sr(1)/10**(-dig))+0.5d0),sr(1))
        para(3*jk-1)=dsign(dint(dabs(sr(2)/10**(-dig))+0.5d0),sr(2))
       enddo
c       write(*,*) 'stcc'
c       write(*,*) (stcc(i,1),i=1,maxm(1))
c       write(*,*) (stcc(i,2),i=1,maxm(2))
       opstk(1)=-1.d05
       do m1=1,maxm(1)
        para(3)=dint(m1/10**(-dig)+0.5d0)
        do m2=1,maxm(2)
         para(6)=dint(m1/10**(-dig)+0.5d0)
         call deslth(para,6,delnth)
```

```
c       write(*,*) 'delnth2', delnth
        stck=stcc(m1,1)+stcc(m2,2)+delnth+d2lg(10**(dabs(dig)))
        if ((opstk(1).eq.(-1.d05)).or.(opstk(1).gt.stck)) then
   opstk(1)=stck
         opstk(2)=m1
         opstk(3)=m2
        else if ((opstk(1).eq.stck).and.(dble(m1+m2).lt.
     &   (opstk(2)+opstk(3)))) then
         opstk(2)=m1
         opstk(3)=m2
        end if
       enddo
      enddo
      return
      end


      subroutine stcmpk(obsn,srk,mk,nk,k,maxn,dig,stck)
      implicit double precision (a-h,o-z)
      integer k, maxn, mk(100), nk(100),nvec(10000)
      dimension obsn(k,maxn),srk(k,2)
      dimension sr(2),para(300)
      external d2lg

c     this subroutine is used to compute the idealized
c     codelength ('stck') of all 'k' samples given the
c     numbers ('mk') of equal-width bins and 'srk' which
c     is the smallest values and the ranges for the k samples,
c     it consists of two parts: one is the stochastic complexity
```

```
c     given 'mk' and 'srk', the other part is the minimum
c     description length  for mk' and 'srk'.
c     'obsn' is the data of k samples, 'nk' are their sample
c     sizes and 'maxn' is the maximum sample size.

      stck=0.d0
      do jj=1,k
       nvec(1)=nk(jj)
       if (mk(jj).gt.1) then
        nvec(1)=0
        do i=1,nk(jj)
         if ((obsn(jj,i).ge.srk(jj,1)).and.(obsn(jj,i).le.
     &     (srk(jj,1)+(1.d0/mk(jj))*s rk(jj,2)))) then
    nvec(1)=nvec(1)+1
         endif
        enddo
        nvec(mk(jj))=0
         do i=1,nk(jj)
          if (obsn(jj,i).gt.(srk(jj,1)+((mk(jj)-1)/dble(
     &       mk(jj)))*srk(jj,2))) then
          nvec(mk(jj))=nvec(mk(jj))+1
         endif
        enddo
        if (mk(jj).gt.2) then
         do j=2,(mk(jj)-1)
          nvec(j)=0
          do i=1,nk(jj)
           if ((obsn(jj,i).gt.(srk(jj,1)+((j-1)/dble(mk(jj)))
     &       *srk(jj,2))).and.(obsn(jj,i).le.(srk(jj,1)+
```

```
     &       (j/dble(mk(jj)))*srk(jj,2)))) then
             nvec(j)=nvec(j)+1
             endif
            enddo
           enddo
          endif
         endif
c         write(*,*) 'nvecpool', (nvec(j),j=1,mk(jj))
         sr(1)=srk(jj,1)
         sr(2)=srk(jj,2)
         call cplxty(sr,nk(jj),mk(jj),nvec,cplx)
         stck=stck+cplx
c         write(*,*) 'stck',stck
         para(3*jj-2)=dsign(dint(dabs(sr(1)/10**(-dig))+0.5d0),sr(1))
         para(3*jj-1)=dsign(dint(dabs(sr(2)/10**(-dig))+0.5d0),sr(2))
         para(3*jj)=dint(mk(jj)/10**(-dig)+0.5d0)
        enddo
        call deslth(para,3*k,delnth)
c        write(*,*) 'delnth', delnth
        stck=stck+delnth+d2lg(10**(dabs(dig)))
        return
        end


        subroutine cplxty(sr,n,m,nvec,cplx)
        implicit double precision (a-h, o-z)
        integer n,m,nvec(10000),nvec1(10000)
        dimension sr(2)
        external d2lg
```

```
c     this subroutine is used to compute the stochastic
c     complexity (cplx) of a set of data Y relative to a
c     set of histogram density functions, given the minimum
c     value of Y (sr(1)), the length of the range of Y
c     (sr(2)), the number of equal-width bins in the histogram
c     density (m) and the number of observations occurring in
c     each equal-width bin (nvec). Here n is the number
c     of the observations.

      nsum=0
      do i=1,n
       nsum=nsum+nvec(i)
      enddo
      if (nsum.ne.n) then
      write(*,*) 'summation of the number in each equal-bin
     & not equal to n'
      else
       dun=n*d2lg(sr(2)/m)
       if (n.eq.1) then
        cplx=dun+d2lg(dfloat(m))
       else
        a1=dun
        do i=m, (n+m-1)
         a1=a1+d2lg(dfloat(i))
        enddo
        a2=0d0
        ncnt=0
        do j=1,m
         if (nvec(j).gt.1) then
```

```
      ncnt=ncnt+1
            nvec1(ncnt)=nvec(j)
           endif
          enddo
          if (ncnt.ne.0) then
           do jj=1,ncnt
      do ii=1,nvec1(jj)
             a2=a2+d2lg(dfloat(ii))
            enddo
           enddo
          endif
          cplx=a1-a2
          endif
         endif
        return
        end


        subroutine deslth(x,mm,delnth)
c       A subroutine to compute the minimum description
c       length (delnth) of a sequence of integers (stored
c       in x),'mm' is the length of 'x'.
        implicit double precision (a-h,o-z)
        integer mm
        dimension x(mm)
        external dlgstr, d2lg
        mplus=0
        sumx=0.d0
        do i=1,mm
```

```
  if (x(i).ge.0.d0) then
   mplus=mplus+1
  endif
  sumx=sumx+dabs(x(i))
enddo
dum=d2lg(2.865064d0)+dlgstr(sumx+1.d0)
if (mm.eq.1) then
 delnth=dum+d2lg(sumx+1.d0)+1.d0
 else if ((mplus.eq.mm).or.(mplus.eq.0)) then
  delnth=dum + d2lg(mm+1.d0)
  do i=1,mm
   delnth=delnth+d2lg(sumx+dfloat(i))
  enddo
  do j=1,(mm-1)
   delnth=delnth-d2lg(dfloat(j))
  enddo
 else
  delnth=dum
  do i=1,mm
   delnth=delnth+d2lg(sumx+dfloat(i))
  enddo
  do j=1, (mm-1)
   delnth=delnth-d2lg(dfloat(j))
  enddo
  do ii=(mplus+1),(mm+1)
   delnth=delnth+d2lg(dfloat(ii))
  enddo
  do jj=1, (mm-mplus)
   delnth=delnth-d2lg(dfloat(jj))
```

```
  enddo
end if
return
end


function dlgstr(x)
implicit double precision (a-h,o-z)
external d2lg
dlgstr=0.d0
dumm=d2lg(x)
do while(dumm .gt. 0.d0)
  dlgstr=dlgstr+dumm
  dumm=d2lg(dumm)
end do
return
end


function d2lg(x)
implicit double precision (a-h, o-z)
temp=2.d0
d2lg=dlog(x)/dlog(temp)
return
end
```

```
function(obs, m, dig, dig.e)
{
#: Splus function "stcmplxty1".
# This function is used to compute the stochastic complexity
# of a set of data relative to the class of histogram
# densities with m equal-width bins. It consists of two
# parts, one is the stochastic complexity given m, minimum
# value of the data (ss), width of the range of the data (r)
# and precision of the data (d); the other part is the
# minimum description length used to describe {s,r,d,m}.

# obs : a vector of observations
# dig: number of decimal digits after the decimal point.
# dig.e: 10^(-dig.e) is the precision set for the
# parameter (ss,r,m)

obs1 <- round(obs, dig)
ran <- range(obs1)
n <- length(obs1)
r <- ran[2] - ran[1]
ss <- ran[1]
para <- round(c(ss/10^( - dig.e), r/10^( - dig.e),
        m/10^( - dig.e)), 0)
nvec <- rep(0, m)
nvec[1] <- n
if(m > 1) {
nvec[1] <- length(obs1[(obs1 >= round(ss, dig)) &
          (obs1 <= round(ss + (1/m) * r, dig))])
```

```
if(m > 1)
for(i in 2:m) {
nvec[i] <- length(obs1[(obs1 > round(ss +
   ((i - 1)/m) * r, dig)) & (obs1
   <= round(ss + (i/m) * r, dig))])
}
}
#print(nvec)
stcmp <- cmplxty1(ss, r, n, m, nvec) + deslen1(para)
 + log(10^(abs( dig.e)), 2)
return(c(stcmp, m))
}


function(ss, r, n, m, nvec)
{
#: Splus function "cmplxty1".
# This function is used to compute the stochastic
# complexity of a set of data Y relative to a set
# of histogram density functions, given the
# minimum value of Y (ss), the length of the range
# of the data (r), the number of equal-width bins
# in the histogram density (m) and the number of
# observations occurring in each equal-width bin
# (nvec). Here n is the number of the observations.

if((length(nvec) != m) | (sum(nvec) != n))
    return("data unmatched")
return(n * (log(r, 2) - log(m, 2)) -
```

```
  (lgamma(m))/log(2) + (lgamma(n + m))/log(2)
  - sum(lgamma(nvec + 1))/log(2))
}


function(x)
{
#: Splus function "deslen1".
# A function to compute the minimum description
# length of a sequence of integers (stored in a
# vector x)

mplus <- length(x[x >= 0])
m <- length(x)
n <- sum(abs(x))
return(log(2.865064, 2) + log.star(n + 1) +
    (1/log(2)) * (lgamma(n + m + 1) - lgamma(n + 1)
    - lgamma(m) + lgamma(m + 2) - lgamma(mplus + 1)
    - lgamma(m - mplus + 1)))
}


function(x)
{
# log.star(x)=log(x,2)+log(log(x,2),2)+
#  log(log(log(x,2),2),2)+... where
# the sum includes all the positive iterates.

dum <- 0
```

```
dum1 <- log(x, 2)
while(dum1 > 0) {
dum <- dum + dum1
dum1 <- log(dum1, 2)
}
return(dum)
}
```

Bibliography

[1] Akaike, H. (1970). Statistical Predictor Identification. *Ann. Inst. Statist. Math.* **22**, 202-217.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symoposium on Information Theory* (B.N. Petrov and F. Czàki, eds.) Akademiai Kiadò, Budapest, 267-281.

[3] Akaike, H. (1974a). A new look at statistical model identification. *IEEE Trans. Automatic Control* **19**, 716-723.

[4] Akaike, H. (1974b). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory,* Ed. B.N. Petrov and F. Csaki, Akademia Kiedo, Budapest, 267-281.

[5] Akaike, H. (1977). On entropy maximization principle. *Applications of statistics.* Ed. P.R. Krishnaiah, North Holland, Amsterdam, 27-41.

[6] Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34**, 122-148.

[7] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc., New York.

[8] Barron, A.R. and Cover, T.M. (1991). Minimum Complexity Density Estimation. *IEEE Trans. on Information Theory* **37**, 1034-1054.

[9] Barron, A.R., Olive, D. and Yang Y. (1992). Asymptotically optimal complexity-based model selection. *presented at IMS Western Regional Meeting (with Biometric Society/WNAR and ASA)*, Corvallis, Oregon, USA, June 1992.

[10] Bentley, J.L. and Yao, A.C. (1976). An almost optimal algorithm for unbounded searching. *Inform. Procrssing Letters* **5**, 82-87.

[11] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second Edition, Springer-Verlag, New York.

[12] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, MA.

[13] Breiman, L.A. and Freedman, D.F. (1983). How many variables should be entered in a regression equation?. *J. Amer. Statist. Assoc.* **78**, 131-136.

[14] Chaitin, G.J. (1975). A theory of program size formally identical to information theory. *J. ACM* **22**, 329-340.

[15] Chow, Y.S. (1960). A Martingale inequality and the law of large numbers. *Proc. Amer. Math. Soc.* **11**, 107-111.

[16] Chow, Y.S. (1967). On a strong law of large numbers for martingales. *Ann. Math. Statist.* **38**, 610-611.

[17] Conover, W.J. (1971). *Practical Nonparametric Statistics.* Wiley, New York.

[18] Davis, A.W. (1977). Asymptotic theory for principal component analysis: Non-normal case. *Austral. J. Statisti.*, **19**, 206-212.

[19] Dawid, A.P. (1984). Present position and potential developments: some personal views. Statistical theory. The prequential approach. *J. Roy. Statist. Soc. A* **47**, 278-292, (with discussion).

[20] Dawid, A.P. (1991a). Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. B* **53**, 79-109, (with discussion).

[21] Dawid, A.P. (1991b). Prequential data analysis. *Issues and Controversies in Statistical Inference* (M. Ghosh and P.K. Pathak, eds.).

[22] Dawid, A.P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A.P. Dawid and A.F.M. Smith, eds), Oxford University Press, 109-125,(with discussion).

[23] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM Algorithm (with discussion). *J. Roy. Statist. Soc. B* **39**, 1-38.

[24] DeVito, C.L. (1978). *Functional Analysis*, Academic Press, New York, 1978.

[25] Devroye, L. (1987) *A Course in Density Estimation*, Birhäuser Boston.

[26] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

[27] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

[28] Elias, P. (1975). Universal codeword sets and representations of the integers. *IEEE Trans. Information Theory* **21**, 194-203.

[29] Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture distributions.* Chapman and Hall, London.

[30] Ferguson, T.S. (1967). *Mathematical Statistics, A Decision Theoretic Approach*, Academic Press, New York and London.

[31] Fienberg, S.E. (1970). The analysis of multidimensional contingency tables. *Ecology* **51**, 419-433.

[32] Freedman, D. A., Diaconis, P. (1981). On the histogram as a density estimator: L^2 theory. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **57**, 453-475.

[33] Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *J. American Stat. Assoc.* **74**, 153-160.

[34] Gerencsér, L. (1989) On Rissanen's predictive stochastic complexity for stationary ARMA processes. McGill Res. Center for Intelligent Machines, TR-CIM-89-5, 1989.

[35] Gerencsér, L. (1992) AR(∞) estimation and nonparametric stochastic complexity. *IEEE Trans. Inf. Theory* **38**, 1768-1778.

[36] Gerencsér, L. and Rissanen, J. (1992) Asymptotics of predictive stochastic complexity. *New Directions in Time Series Analysis. Part II, Proc. 1990 IMA Workshop, IMA Volumes in Mathematics and Its Applications.* P.E. Caines, J. Geweke, and M. Taqqu, Eds. New York: Springer, 1992.

[37] Good, I.J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* University of Minnesota Press, Minneapolis.

[38] Gunst, R.F. and Mason, R.L. (1980). *Regression Analysis and its Application.* Dekker, New York.

[39] Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

[40] Hall, P. and Hannan, E.J. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika* **75**, 705-714.

[41] Hannan, E.J., McDougall, A.J. and Poskitt, D.S. (1989). Recursive estimation of autoregressions. *J. Roy. Stat. Soc. B* **51**, 217-233.

[42] Hazewinkel, M. et al. (1991). *Encyclopaedia of Mathematics*, vol. 7, Kluwer Academic Publishers, Dordrecht, 1991.

[43] Hemerly, E.M. and Davis, M.H.A. (1989). Strong consistency of the predictive least squares criterion for order determination of autoregressive precesses. *Ann. Statist.* **17**, 941-946.

[44] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks.* Wiley, New, York.

[45] Hjorth, U. (1982). Model selection and forward validation. em Scand. J. Stat. **9**, 95-105.

[46] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417-441, 498-520.

[47] Jain, R. (1983). *Probabilistic weather forecasting.* MSc. Dissertation, Department of Statistical Science, University College London.

[48] Jaynes, E. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106**, 620, **108**, 171.

[49] Jaynes, E.T. (1978). Where do we stand on maximum entropy? *E.T. Jaynes: papers on probability, statistics, and statistical physics*, edited by R. D. Rosenkrantz, Dordrecht, Holland, pp. 210-314, 1978.

[50] Jaynes, E. (1982). On the rationale of maximum entropy methods. *Proc. of IEEE, Special Issue on Spectral Estimation.* S. Haykin, editor, **70**, 939-952.

[51] Jaynes, E. (1985). Where do we go from here? *Maximum-Entropy and Bayesian Methods in Inverse Problem* eds. C.R. Smith and W.T. Grandy, D. Reidel Publ. Co., 21-58.

[52] Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information. em Problems of Information Transmission **1**, 4-7.

[53] Kshirsagar, A.M. (1972). *Multivariate analysis*, Marcel Dekker, Inc., New York.

[54] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

[55] Lehmann, E.L. (1986a). *Theory of Point Estimate.* 2nd Edition, Wiley, New York.

[56] Lehmann, E.L. (1986b). *Testing Statistical Hypotheses.* 2nd Edition, Wiley, New York.

[57] Leung-Yan-Cheong, S.K. and Cover, T. (1978). Some equivalences between Shannon entropy and Kolmogorov complexity. *IEEE Trans. Inf. Theory* **24**, 331-338.

[58] Lock, R. (1992). PRO Football Scores, StatLib.

[59] Lopuhaa, H.P. and Rousseeuw, P.J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19**, 229-248.

[60] Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[61] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models,* Chapman and Hall, London.

[62] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory,* John Wiley & Sons, Inc., New York.

[63] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-765.

[64] O'Hagan, A. (1994). Fractional Bayes Factors for Model Comparison. To appear in *J. Roy. Stat. Soc. B.*

[65] Randles, R.H. and Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York.

[66] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications, 2nd Edition*, John Wiley & Sons, Inc., New York.

[67] Rissanen, J. (1976). Generalized Kraft-Inequality and arithmetric coding. *IBM J. Res. Devel.* **20**, 198-203.

[68] Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465-471.

[69] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416-431.

[70] Rissanen, J. (1986a). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080-1100.

[71] Rissanen, J. (1986b). Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory* **32** 526-532.

[72] Rissanen, J. (1986c) A predictive least squares principle. *IMA J. Math. Contr. Inform.* **3**, 211-222.

[73] Rissanen, J. (1986d) Order estimation by accumulated prediction errors. In *Essays in Times Series and Allied Processes* (J. Gani and M.B. Priestley, eds.) 55-61, Applied Probability Trust, Scheffield, England.

[74] Rissanen, J. (1987). Stochastic complexity. *J. Roy. Stat. Soc. B* **49**, 223-265(with discussions).

[75] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Co. Pte. Ltd. , Singapore.

[76] Rissanen, J. (1994a). Fisher information and stochastic complexity. to appear in *IEEE Trans. Inform. Theory.*

[77] Rissanen, J. (1994b). Information theory and neural nets. *Mathematical Perspectives on Neural Networks* (eds. P. Smolensky, M. Mozer and D. Rumelhart), Laurence Erlbaum Associates (to appear).

[78] Rissanen, J. and Langdon, Jr. G.G. (1981). Universal modeling and coding. *IEEE Trans. on Inform. Theory* bf 27, 12-23.

[79] Rissanen, J. and Mohiuddin, K. (1989). A multiplication-free multialphabet arithmetic code. *IEEE Trans. on Communications* **37**, 93-98.

[80] Rissanen, J. and Ristad, E.S.(1992). Unsupervised classification with stochastic complexity. *presented at the First US/Japan Conf. on the Frontiers of Statistical Modeling: an Informational Approach,* Knoxville, Tennessee, USA, May. 1992.

[81] Rissanen, J., Speed, T.P. and Yu, B. (1992). Density Estimation by Stochastic Complexity. *IEEE Trans. Information Theory* **38**, 315-323.

[82] Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3**, 1-14.

[83] Rousseeuw, P.J. (1991). A diagnostic plot for regression outliers and leverage points. *Computational Statistics and Data Analysis* **11**, 127-129.

[84] Rousseeuw, P.J. and van Zomeren, B.C. (1990), Unmasking Multivariate Outliers and leverage points (with discussion). *J. Amer. Statist. Assoc.* **85**, p. 633-651.

[85] Schoener, T.W. (1970). Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology* **51**, 408-418.

[86] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

[87] Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **47**, 143-157.

[88] Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

[89] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

[90] Shibata, R. (1983a). Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math.* **35**, 415-423.

[91] Shibata, R. (1983b). A theoretical view of the use of AIC. *Times Series Analysis: Theory and Practice 4* (ed O. D. Anderson), 237-244, Elsevier, Amsterdam.

[92] Shiryayev, A. N. (1984) *Probability*, New York: Springer-Verlag.

[93] Solomonoff, R.J. (1964). A formal theory of inductive inference. Part I, *Information and Control* **7**, 1-22; Part II, *Information and Control* **7**, 224-254.

[94] Solomonoff, R.J. (1978). Complexity-based induction systems: comparison and convergence theorems. *IEEE Trans. Information Theory* **24**, 422-432.

[95] Speed, T.P. and Yu, B. (1993). Model selection and prediction: normal regression. *Ann. Inst. Statist. Math.* **45**, 35-54.

[96] Stone, C. J. (1985). An asymptotic optimal histogram selection rule. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (Le Cam, L. M., Ohshen, R. A., eds.) vol II, 513-520. Belmont, CA: Wadsworth.

[97] Stone, M. (1974). Cross-validation choice and assessment of statistical predictions *J. Roy. Stat. Soc. B* **36**, 111-147.

[98] Stone, M. (1977a). Asymptotics for and against Cross-Validation. *Biometrika* **64**, 29-35.

[99] Stone, M. (1977b). An asymptotic equivalence of choice of model by Cross-Validation and Akaike's criterion. *J. Roy. Stat. Soc. B* **39**, 44-47.

[100] Stout, W.F. (1974). *Almost Sure Convergence,* Academic Press, New York.

[101] Valentine, F.A. (1964). *Convex Sets*, McGraw-Hill, New York.

[102] Wald, A.(1971). *Statistical Decision Functions.* Second Edition, Chelsea Publ. Co., New York, 1971.

[103] Wallace, C.S., Freeman, P.R. (1987). Estimation and Inference by compact coding. *J. Roy. Statist. Soc. B* **9**, 240-251 and 252-265 (discussions).

[104] Waternaux, C.M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. *Biometrika* **63**, 3, 639-45.

[105] Wax, M. (1988). Order selection for AR models by predictive least squares. *IEEE Trans. Acoust. Speech Signal Process.* **36**, 581-588.

[106] Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439-447.

[107] Wei, C.Z. (1992). On the predictive least squares principle. *Ann. Statist. 20*, 1-42.

[108] Wolfe, J.H. (1970). Pattern Clustering by Multivariate mixture analysis. *Multivar. Behav. Res.* **5**, 329-350.

[109] Yu, B. and Speed, T.P. (1992). Data compression and histograms. *Probab. Theory Relat. Fields* **92**, 195-229.

[110] Zhvonkin, A.K. and Levin, L.A. (1970). The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys* **25**, 83-124.