ON NEW DEVELOPMENTS IN STATISTICAL INFERENCE FOR MEASURES OF DIVERGENCE

By

Kyriacos Mattheou

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
UNIVERSITY OF CYPRUS
NICOSIA, CYPRUS
DECEMBER 2007

UNIVERSITY OF CYPRUS
DEPARTMENT OF
MATHEMATICS AND STATISTICS

The undersigned hereby certify that they have read and recommend to the senate for acceptance a thesis entitled "**On new developments in statistical inference for measures of divergence**" by **Kyriacos Mattheou** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: December 2007

Examining Committee:

Tasos Christofides, Chairman
Professor, University of Cyprus

Alex Karagrigoriou, Research Advisor
Associate Professor, University of Cyprus

Takis Papaioannou
Professor, University of Pireaus
and University of Ioannina

Leandro Pardo
Professor, Complutense University of Madrid

Ilia Vonta
Assistant Professor, University of Cyprus

UNIVERSITY OF CYPRUS

Date: **December 2007**

Author: **Kyriacos Mattheou**

Title: **On new developments in statistical inference for measures of divergence**

Department: **Mathematics and Statistics**

Degree: **Ph.D.** Convocation: **July** Year: **2008**

Permission is herewith granted to University of Cyprus to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

To Nikoletta, Konstantina and Eleni

Table of Contents

List of Tables

List of Figures

Abstract

An issue of fundamental importance in Probability and Statistics is the investigation of Information Measures. These measures are classified in different categories and measure the quantity of information contained in the data with respect to a parameter θ, the divergence between two populations or functions, the information we get after the execution of an experiment and other important information according to the application they are used for.

A literature review on the measures of information, classified in four main categories namely divergence - type, entropy - type, Fisher - type and Bayesian - type is provided. Special attention is given to the divergence - type measures.

In this work we first propose and investigate a general family of measures of divergence which is based on the BHHJ measure of divergence of Basu, Harris, Hjort, and Jones (1998). A number of main properties of the family, such as the nonnegativity property, the continuity property, the invariance property, the symmetric property, the limiting property, the order preserving property and the quadratic convergence are discussed.

Since measures of divergence are used as indices of similarity or dissimilarity between populations, they can be used to develop new model selection criteria. Applying the same methodology used for the well known Akaike Information Criterion (AIC), a new model selection criterion called *Divergence Information Criterion (DIC)* is proposed as an approximately unbiased estimator of the expected overall BHHJ discrepancy (divergence).

Then, we focus on the investigation of the discrete form of the measure. We provide the distributional properties of the estimator of this general family of BHHJ measures of divergence and we propose a test statistic based on this family of measures for goodness of fit tests for multinomial distributions.

Finally, a number of simulations are performed to check the appropriateness of

both the proposed model selection criterion and the test statistic for goodness of fit. The simulations for the model selection criterion compare the performance of DIC, with other well known criteria such as the standard BIC and AIC and also some variants of them. The simulations for the goodness of fit test involve the new test statistic based on the BHHJ measure, and the tests based on the Kullback - Leibler, Kagan, Matusita, and Cressie and Read measures.

Περίληψη

Ένα ζήτημα το οποίο θεωρείται πολύ σημαντικό έως θεμελιώδες στη θεωρία πιθανοτήτων και στη στατιστική θεωρία είναι η μελέτη των Μέτρων Πληροφορίας. Τα Μέτρα Πληροφορίας κατηγοριοποιούνται σε διάφορες κλάσεις και μετρούν την ποσότητα της πληροφορίας που περιέχεται στα δεδομένα σε σχέση με μια άγνωστη παράμετρο, την απόκλιση (απόσταση) μεταξύ δύο πληθυσμών ή συναρτήσεων, την πληροφορία που εξάγεται μετά την εκτέλεση ενός πειράματος και άλλων μορφών σημαντική πληροφορία, σύμφωνα βέβαια με την εφαρμογή στην οποία τα συναντούμε ή τα χρησιμοποιούμε.

Στην παρούσα διατριβή γίνεται αρχικά μια βιβλιογραφική ανασκόπηση που αφορά τα Μέτρα Πληροφορίας και η οποία περιλαμβάνει μια ταξινόμηση των μέτρων αυτών σε τέσσερις κύριες κατηγορίες ως εξής: μέτρα τύπου απόκλισης, μέτρα τύπου εντροπίας, μέτρα τύπου Fisher και μέτρα τύπου Bayes. Ιδιαίτερη βαρύτητα δίνεται στα μέτρα απόκλισης.

Στη συνέχεια προτείνεται μια νέα γενικευμένη οικογένεια μέτρων απόκλισης, η οποία βασίζεται στο μέτρο απόκλισης BHHJ, το οποίο προτάθηκε από τους Basu, Harris, Hjort και Jones (1998). Για την οικογένεια αυτή, αποδεικνύονται οι κύριες ιδιότητές που αφορούν τη μη αρνητικότητα, τη συνέχεια, το αναλλοίωτο, τη συμμετρία, την ασυμπτωτική συμπεριφορά, τη διατήρηση της διάταξης και την τετραγωνική σύγκλιση.

Τα μέτρα απόκλισης χρησιμοποιούνται ως ενδείξεις ομοιότητας η ανομοιότητας μεταξύ πληθυσμών. Επομένως είναι δυνατόν να χρησιμοποιηθούν μεταξύ άλλων και για την κατασκευή νέων κριτηρίων επιλογής μοντέλων. Εφαρμόζοντας ανάλογη μεθοδολογία με αυτήν που χρησιμοποιήθηκε για την κατασκευή του γνωστού κριτηρίου του Akaike (Akaike Information Criterion, AIC, Akaike, 1973) προτείνεται ένα νέο κριτήριο επιλογής μοντέλων, το Divergence Information Criterion (DIC) που προκύπτει ως μια αμερόληπτη εκτιμήτρια της αναμενόμενης ολικής BHHJ απόκλισης. Επίσης προσδιορίζεται το κάτω φράγμα του μέσου τετραγωνικού σφάλματος πρόβλεψης.

Ακολούθως, η διατριβή επικεντρώνεται στη διερεύνηση της διακριτής μορφής της

νέας γενικευμένης οικογένειας μέτρων απόκλισης BHHJ και αποδεικνύονται οι ιδιότητες της κατανομής της εκτιμήτριάς της. Επίσης προτείνεται μια νέα στατιστική συνάρτηση για ελέγχους υποθέσεων καλής προσαρμογής σε πολυωνυμικούς πληθυσμούς και αποδεικνύεται η ασυμπτωτική κατανομή της κάτω από την μηδενική υπόθεση όπως και κάτω από την εναλλακτική υπόθεση της συνάφειας (contiguous alternative).

Τέλος παρουσιάζονται μια σειρά εφαρμογών για διερεύνηση της καταλληλότητας του κριτηρίου επιλογής μοντέλων DIC καθώς και της στατιστικής συνάρτησης για τους ελέγχους καλής προσαρμογής. Στις εφαρμογές για το κριτήριο επιλογής μοντέλων συγκρίνεται η απόδοση του DIC, με άλλα γνωστά κριτήρια, όπως τα Bayesian Information Criterion (BIC, Scharz, 1977) και AIC, καθώς και κάποιες ειδικές μορφές τους. Στις εφαρμογές για τους ελέγχους καλής προσαρμογής γίνεται σύγκριση της προτεινόμενης στατιστικής συνάρτησης η οποία βασίζεται στο μέτρο BHHJ και των ελέγχων που βασίζονται στα μέτρα Kullback-Leibler, Kagan ('Ελεγχος Καλής Προσαρμογής του Pearson), Matusita και Cressie and Read.

Acknowledgements

I would like to thank Dr. Alex Karagrigoriou, my supervisor, for his many suggestions and constant support during this research.

I am also thankful to Dr. Takis Papaioannou, Dr. Konstantinos Zografos, Dr. Ilia Vonta, Dr. Panagiotis Mantalos, Dr. Sangyeol Lee, Dr. Tasos Christofides, Dr. Leandro Pardo, and Dr. Konstantinos Fokianos for their kind advice and constructive suggestions.

Finally, I am grateful to my wife Eleni for her patience and support. Without her this work would never have come into existence.

Introduction

The divergence (or discrepancy) measures are used as indices of similarity or dissimilarity between populations. They are also used to measure the distance or the discrepancy between two functions or two populations. Finally they are used either to measure mutual information concerning two variables or to construct model selection criteria.

Measures of divergence between two probability distributions have a very long history. One could consider as pioneers in this field the famous Mathematicians and Statisticians of the 20th century, Pearson, Mahalanobis, Lévy and Kolmogorov. In our days the most popular measure of divergence is considered the Kullback-Leibler measure of divergence introduced in the 50's. A well known family of measures is the φ-divergence known also as Csiszar's measure of information which was introduced and investigated independently by Csiszár (1963) and Ali and Silvey (1966). For various functions for φ the measure takes different forms. Members of this family are among others, the Kullback-Leibler measure as well as Pearson's X^2 divergence measure also known as Kagan's divergence measure.

A unified analysis has been provided by Cressie and Read (1984, 1988) who introduced for both the continuous and the discrete case a family of measures of divergence known as power divergence family of statistics that depends on a parameter λ and is used for goodness-of-fit tests for multinomial distributions. The Cressie and Read family includes among others the well known Pearson's X^2 divergence measure and for multinomial models the loglikelihood ratio statistic. It should be noted that for the appropriate limit of λ to 0 the above measure becomes the Kullback-Leibler measure.

A new measure of divergence known as the BHHJ divergence measure, was recently introduced by Basu et *al.* (1998). The measure depends on an index a which controls the trade-off between robustness and efficiency when the measure is used as an estimating criterion for robust parameter estimation. Basu et *al.* (1998) showed that values of a close to zero provide parameter estimators with good robust features without significant loss in terms of efficiency. Note that for the appropriate limit of a to 0 the measure reduces to the Kullback-Leibler measure.

As it was mentioned earlier measures of divergence can also be used in model selection. Since some measures of divergence have been proposed as distinguishability indices between two distributions which are far from each other or from two distributions which are close, they can be used for the construction of model selection criteria. A model selection criterion can be considered as an approximately unbiased estimator of the expected overall discrepancy, a nonnegative quantity which measures the *distance* between the true unknown model and a fitted approximating model belonging to a class of candidate models. If the value of the criterion is small for a specific member of the candidate class then the corresponding approximated model is good. The Kullback-Leibler divergence was the measure used by Akaike (1973) to develop the Akaike Information Criterion (AIC).

In this work, we focus on the BHHJ measure of divergence and we propose a general class of continuous BHHJ divergence measures that includes the BHHJ divergence measure of Basu et *al.* (1998) as well as a general class of discrete measures of divergence which could be viewed as the discrete version of the above continuous BHHJ class and could be used for goodness of fit tests. The continuity and discrete character of the new class will be explained in the last section of Chapter 1. This new class of measures is fully investigated in this thesis, by

- establishing a number of properties (Chapter 2),
- developing a new model selection criterion, the Divergence Information Criterion (DIC) (Chapter 3) and
- introducing a new class of test statistics for performing goodness of fit tests (simple null hypothesis) for multinomial populations (Chapter 4).

Simulation results are provided in Chapter 5 for testing the appropriateness of the proposed criterion as well as the test statistics. Chapter 1 is devoted to a Literature Review on measures of divergence and the presentation of the new BHHJ class of measures. The work is concluded with a Discussion and a Future Research plan.

Chapter 1

Measures of Information-Literature Review

Information Theory in Probability and Statistics has a very long history and it is of fundamental importance. There are many approaches and definitions, for Information in Statistics, from different authors and from different aspects.

'*While information is a basic and fundamental concept in statistics there is no universal agreement on how to define and measure it in a unique way*' (Papaioannou, 2001). There have been several statements made over the years. For more details on the variety of views see the review articles by Kendall (1973), Csiszár (1977), Papaioannou (1985), Aczel (1986), Soofi (1994), Pardo (1999), Kullback (1959), Papaioannou and Kempthone (1971) and Ferentinos and Papaioannou (1981).

Although not universally accepted, there is a classification of measures of information in four categories namely,

- Divergence - type,
- Entropy - type,
- Fisher - type and
- Bayesian - type.

Representative measures in each category are

- the Kullback-Leibler divergence (1951),
- the Shannon's entropy (1948),

- the Fisher information measure (1925) and

- the Lindley's measure of information (1956)

correspondingly.

In the present literature review, important measures of information that play a significant role in statistical inference with numerous applications are presented. Special attention is given to measures of divergence.

Two classical measures that illustrate the fundamental importance of the measures of information are the Kolmogorov Distance and the Lévy Distance introduced in the 30's and 20's respectively.

Let μ and ν probability measures on $\mathbb{R}$ with associate distribution functions F_1 and F_2. Kolmogorov Distance (Kolmogorov, 1933) is defined as:

$$K\left(F_1, F_2\right) = \sup_{x \in \mathbb{R}} \left|F_1\left(x\right) - F_2\left(x\right)\right|.$$

An important implementation of the Kolmogorov distance is the well known Glivenko-Cantelli Theorem which states that the Empirical Distribution Function

$$F_n\left(x\right) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}\left(x_i\right)$$

where $I_{(-\infty,x]}\left(x_i\right) = 1$ if $x_i < x$ and 0 otherwise, is uniformly strongly consistent for the true Distribution Function F in the sense that:

$$\lim_{n \to \infty} P\left\{K\left(F_n, F\right) > \varepsilon\right\} = 0, \forall\ \varepsilon > 0.$$

On the other hand, Lévy Distance (Lévy, 1925) between two distribution functions F_1 and F_2 is defined as:

$$L\left(F_1, F_2\right) = \inf\left\{\varepsilon > 0 : F_1(x-\varepsilon) - \varepsilon \leq F_2(x) \leq F_1(x+\varepsilon) + \varepsilon, \forall x \in \mathbb{R}\right\}.$$

Note that this distance is not easy to compute and convergence in Lévy Distance means weak convergence for distribution functions in $\mathbb{R}$. For the relations between the Kolmogorov and Lévy Distances see Gibbs and Su (2002). Since μ and ν are measures on $\mathbb{R}$, it is customary to view the Lévy as well as the Kolmogorov distance as measures of distance (divergence) between the corresponding distribution functions F_1 and F_2.

1.1 Measures of Divergence

A measure of divergence is used as a way to evaluate the distance or divergence between any two populations or functions. Let f_1 and f_2 be two probability density functions which may depend or not on an unknown parameter of fixed finite dimension. The most well known measure of (directed) divergence is the Kullback-Leibler divergence which is given by

$$I_X^{KL}(f_1, f_2) = \int f_1 \log(f_1/f_2) d\mu = E_{f_1}\left[\log\left(\frac{f_1}{f_2}\right)\right],$$

for a measure μ which, for the continuous case, is the Lebesgue measure, and a random variable X with absolutely continuous distribution. This means that for a density f with probability distribution P, associated with the continuous random variable X,

$$f(x) = \frac{dP}{d\mu}(x), \text{ where } \mu \text{ is the Lebesque measure.}$$

The above notation covers not only the continuous case but also a discrete setting where the measure μ is a counting measure. Indeed, for the discrete case, the divergence is meaningful for the probability mass functions f_1 and f_2 whose support is a subset of the support S_μ, finite or countable, of the counting measure μ that satisfies $\mu(x) = 1$ for $x \in S_\mu$, and 0 otherwise. In this setting for a probability mass function f with probability distribution P, we have

$$f(x) = \frac{dP}{d\mu}(x) = P(X = x), \text{ where } \mu \text{ is the counting measure}$$

and X a discrete random variable.

So, for the above divergence and for the subsequent ones consider that, if k is a measurable function, the expectation of $k(X)$ is given by:

$$E_f[k(X)] = \begin{cases} \int k(x) f(x) dx, & \text{if } \mu \text{ is the Lebesgue measure} \\ \sum\limits_{x \in S_\mu} k(x) f(x), & \text{if } \mu \text{ is the counting measure} \end{cases}.$$

If f_1 is the density of $X = (U, V)$ and f_2 is the product of the marginal densities of U and V, I_X^{KL} is the well known mutual or relative information in coding theory. The Kullback-Leibler divergence is also looked upon as discriminatory information.

Jeffreys (1946) defined the symmetric divergence:

$$I_X^J(f_1, f_2) = I_X^{KL}(f_1, f_2) + I_X^{KL}(f_2, f_1)$$

Observe that Jeffreys' measure as opposed to the Kullback-Leibler measure is a symmetric measure.

As generalizations of the Kullback-Leibler measure, the additive and non-additive directed divergences of order α were introduced in the 60's and the 70's (Rényi, 1961, Csiszár, 1963 and Rathie and Kannappan, 1972). The so called order α information measure of Rényi (1961) is given by

$$\begin{aligned} I_X^{R,\alpha}(f_1, f_2) &= \frac{1}{\alpha - 1} \log \int f_1^{\alpha} f_2^{1-\alpha} d\mu \\ &= \frac{1}{\alpha - 1} \log E_{f_1}\left[\left(\frac{f_1(X)}{f_2(X)}\right)^{\alpha-1}\right], \quad \alpha > 0, \quad \alpha \neq 1. \end{aligned}$$

It should be noted that for $\alpha \uparrow 1$ (limit by the right) the above measure reduces to the Kullback-Leibler divergence. Observe also that for $\alpha = 1/2$ Rényi's measure becomes the well known Bhattacharyya measure.

An extension of $I_X^{R,\alpha}(f_1, f_2)$ was given by Liese and Vajda (1987), for all $\alpha \neq 0, 1$:

$$\begin{aligned} I_X^{R_{lv},\alpha}(f_1, f_2) &= \frac{1}{\alpha(\alpha - 1)} \log \int f_1^{\alpha} f_2^{1-\alpha} d\mu \\ &= \frac{1}{\alpha(\alpha - 1)} \log E_{f_1}\left[\left(\frac{f_1(X)}{f_2(X)}\right)^{\alpha-1}\right], \quad \alpha \neq 0, 1. \end{aligned}$$

The cases $\alpha = 0, 1$ are defined by continuity:

$$I_X^{R_{lv},1}(f_1, f_2) = \lim_{\alpha \uparrow 1} I_X^{R_{lv},\alpha}(f_1, f_2) = I_X^{KL}(f_1, f_2)$$

and

$$I_X^{R_{lv},0}(f_1, f_2) = \lim_{\alpha \downarrow 0} I_X^{R_{lv},\alpha}(f_1, f_2) = I_X^{KL}(f_2, f_1).$$

The second limit given as $I_X^{KL}(f_2, f_1)$ is also known as the (Mean) Discrimination Information for discriminating f_2 from f_1.

Furthermore, the Matusita measure [Matusita, 1951] given by

$$I_X^M(f_1, f_2) = \int (\sqrt{f_1} - \sqrt{f_2})^2 d\mu$$

is the square of the well known Hellinger distance. Note that Matusita (1964) generalized the above measure for any $0 < a < 1$ (see Table 1.1).

Another measure of divergence is the measure of Kagan (1963) which is known as Pearson's X^2 and is given by

$$I_X^{Ka}(f_1, f_2) = \frac{1}{2}\int (1 - f_1/f_2)^2 f_2 d\mu.$$

Csiszár's measure of information [Csiszár (1963), Ali and Silvey, 1966] is a general divergence-type measure, known also as φ-divergence based on a convex function φ and defined by

$$I_X^{C,\varphi}(f_1, f_2) = \int \varphi(f_1/f_2) f_2 d\mu = E_{f_2}\left[\varphi\left(\frac{f_1}{f_2}\right)\right], \ \varphi \in \Phi^*$$

where Φ^* is the class of all convex functions φ on $[0, \infty)$ such that $\varphi(1) = 0$ and $\varphi''(1) \neq 0$. In the expression of $I_X^{C,\varphi}(f_1, f_2)$ we shall assume the conventions

$$0\varphi(0/0) = 0 \quad \text{and} \quad 0\varphi(u/0) = \lim_{u\to\infty} \varphi(u)/u, \text{ for } u > 0.$$

Remark 1.1.1. *[Pardo (2006)]. If $\varphi \in \Phi^*$ is differentiable at $x = 1$, then the function*

$$\psi(x) \equiv \varphi(x) - \varphi'(1)(x - 1)$$

also belongs to Φ^* and has the additional property that $\psi'(1) = 0$. This property, together with convexity, implies that $\psi(x) \geq 0$, for any $x \geq 0$. Further,

$$\begin{aligned} I_X^{C,\psi}(f_1, f_2) &= \int f_2 \ \left[\phi\left(\frac{f_1}{f_2}\right) - \phi'(1)\left(\frac{f_1}{f_2} - 1\right)\right] d\mu \\ &= \int f_2 \varphi\left(\frac{f_1}{f_2}\right) d\mu \\ &= I_X^{C,\varphi}(f_1, f_2). \end{aligned}$$

Since the two divergence measures coincide, we can consider the set Φ^* to be equivalent to the set

$$\Phi \equiv \Phi^* \cap \{\varphi : \varphi'(1) = 0\}.$$

Observe that Csiszár's measure reduces to Kullback-Leibler divergence if

$$\varphi(u) = u \log u.$$

If $\varphi(u) = (1/2)(1 - u)^2$ or $\varphi(u) = (1 - \sqrt{u})^2$ Csiszár's measure yields the Kagan's and the square of Matusita's divergence respectively.

Table 1.1: Csiszár's Measures of Divergence

φ-function	Divergence
$x \log x - x + 1$ or $x \log x$	Kullback-Leibler (1959)
$-\log x + x - 1$ or $-\log x$	(Mean) Discrimination Information
$(x-1)\log x$	Jeffreys (1946)
$\frac{1}{2}(x-1)^2$	Pearson (1900), Kagan (1963)
$\frac{(x-1)^2}{(x+1)^2}$	Balakrishnan and Saghvi (1968)
$\frac{-x^s+s(x-1)+1}{1-s},\ s \neq 1$	Rathie and Kannappan (1972)
$\frac{1+x}{2} - \left(\frac{1+x^{-r}}{2}\right)^{-1/r},\ r > 0$	Harmonic mean (Mathai and Rathie (1975))
$\frac{(1-x)^2}{2(a+(1-a)x)},\ 0 \leq a \leq 1$	Rukhin (1994)
$\frac{ax\log x-(ax+1-a)\log(ax+1-a)}{a(1-a)},\ a \neq 0,1$	Lin (1991)
$\frac{x^{\lambda+1}-x-\lambda(x-1)}{\lambda(\lambda+1)},\ \lambda \neq 0,-1$	Cressie and Read (1984)
$\lvert 1-x^a\rvert^{1/a}, 0 < a < 1$	Matusita (1964)
$\lvert 1-x\rvert^a,\ a \geq 1$	χ^2 - divergence of order a (Vajda, 1973) Total Variation if $a = 1$ (Saks, 1937)

More examples of φ-functions and the measures we obtain based on these functions are given in Table 1.1 (reproduced from Pardo, 2006).

A well known generalization of measures of divergence is the family of power divergences introduced independently by Cressie and Read (1984) and Liese and Vajda (1987) which is given by

$$I_X^{CR}(f_1,f_2)=\frac{1}{\lambda(\lambda+1)}\int f_1\left[\left(\frac{f_1}{f_2}\right)^{\lambda}-1\right]d\mu,\ \lambda\in R,$$

where for $\lambda=0,-1$ is defined by continuity. Note that the Kullback-Leibler divergence is obtained for $\lambda\downarrow-1$ or $\lambda\uparrow 0$. Note also that as it can be seen in Table 1.1 this divergence is a member of the Csiszár's family of measures.

Although most of the known measures belong to the family of the Csiszár's family of measures there are measures that do not fit into this family. The gap has been fulfilled by a generalization of Csiszár's φ-divergence known as (h,φ) divergence measure. This new family which has been proposed by Menéndez et *al.* (1995) involves an additional differentiable increasing real function h with $h(0)=0$, $h'(0)>0$:

$$I_X^{h,C,\varphi}(f_1,f_2)=h\left(I_X^{C,\varphi}(f_1,f_2)\right).$$

This family of measures has been extensively investigated although the use of two functions (φ and h) increases both its complexity and its applicability.

Some measures included in this general family are Rényi's (Rényi, 1961) and the extension of Rényi's measure (Liese and Vajda, 1987) which were mentioned earlier, Sharma-Mittal's measure (Sharma and Mittal, 1977) given by

$$\begin{aligned}I_X^{s,\alpha}(f_1,f_2)&=\frac{1}{s-1}\left(\left(\int f_1^{\alpha}f_2^{1-\alpha}d\mu\right)^{\frac{s-1}{\alpha-1}}-1\right)\\&=\frac{1}{s-1}\left(\left(E_{f_1}\left[\left(\frac{f_1}{f_2}\right)^{\alpha-1}\right]\right)^{\frac{s-1}{\alpha-1}}-1\right),\end{aligned}$$

for $\alpha,s\neq1$ or

$$\begin{aligned}I_X^{s,1}(f_1,f_2)&=\frac{1}{s-1}\left(\exp\left((s-1)\int f_1\log\frac{f_1}{f_2}d\mu\right)-1\right)\\&=\frac{1}{s-1}\left(\exp\left((s-1)E_{f_1}\left[\log\left(\frac{f_1}{f_2}\right)\right]\right)-1\right),\end{aligned}$$

Table 1.2: (h, φ) Measures of Divergence

Divergence	$h(x)$	$\varphi(x)$
Rényi	$\frac{1}{\alpha(\alpha-1)} \log\left(\alpha\left(\alpha-1\right)x+1\right),\ \alpha \neq 0,1$	$\frac{x^{\alpha}-\alpha(x-1)-1}{\alpha(\alpha-1)},\ \alpha \neq 0,1$
Sharma-Mittal	$\frac{1}{s-1}\left(\left(1+\alpha\left(\alpha-1\right)x\right)^{\frac{s-1}{\alpha-1}}-1\right),\ s,\alpha \neq 0,1$	$\frac{x^{\alpha}-\alpha(x-1)-1}{\alpha(\alpha-1)},\ \alpha \neq 0,1$
Bhattacharyya	$-\log\left(-x+1\right)$	$-x^{1/2}+\frac{1}{2}\left(x+1\right)$

for $s \neq 1$ and Bhattacharyya's measure (Bhattacharyya, 1943) given by

$$I_X^{Bh}\left(f_1, f_2\right) = -\log \int \sqrt{f_1 f_2} d\mu.$$

The above measures are summarized in Table 1.2 (Pardo, 2006). Observe that $4I_X^{Bh} \equiv I_X^{R_{lv,1/2}}$.

1.2 Entropy - Type Measures-Diversities

For historical reasons the representative measure of this category is considered to be Shannon's Entropy (1948) given by

$$I^S\left(X\right) \equiv I^S\left(f\right) = -\int f \log f d\mu = E_f\left[-\log f\right],$$

where X is a random variable with density function $f(x)$.

The word diversity quite often means "variety", referring to a large number (a variety) of different types of the same thing. In a given ecosystem, the variation of life forms is known as biodiversity and is often used as a measure of the health of biological systems. In such cases it is often important to have available a tool to measure how much diversity (variety) there is.

Shannon's entropy was introduced and used during the second World War, in Communication Engineering. Shannon derived the discrete version of $I^S\left(f\right)$ where f a probability mass function and named it *entropy* because of its similarity with thermodynamics entropy. The continuous version was defined by analogy. For a finite number of points, Shannon's entropy measures the expected information of a signal transferred without noise from a source X with density $f(x)$ and is related to

Kullback-Leibler divergence through the following expression:

$$I^S(f) = I^S(h) - I_X^{KL}(f,h)$$

where h is the density of the uniform distribution.

The second most popular entropy measure in discrete settings is Gini-Simpson Index (Gini, 1912, Simpson, 1949). Let $P = (p_1, p_2, \ldots, p_m)$ be a discrete finite probability distribution. Then the discrete version of Gini-Simpson Index is given by:

$$I^{GS}(P) = 1 - \sum_{i=1}^{m} p_i^2.$$

This measure was investigated among others by Agresti and Agresti (1978), Patil and Taille (1982) and Rao (1982).

Many generalizations of Shannon Entropy were hereupon introduced. Rényi's (1961), given by

$$I^{R,a}(f) = \frac{1}{a-1} \log E_f [f]^{a-1}, \ a > 0, \ a \neq 1$$

and Liese and Vajda's (1987) extension of Rényi's Entropy, given by

$$I^{R_{lv},a}(f) = \frac{1}{a(a-1)} \log E_f [f]^{a-1}, \ a \neq 0, 1.$$

Note that for $a \to 1$ and $a \to 0$ we get

$$\lim_{a \to 1} I^{R_{lv},a}(f) = I^S(f)$$

and

$$\lim_{a \to 0} I^{R_{lv},a}(f) = \int \log f d\mu.$$

For more about entropy measures the reader is referred to Mathai and Rathie (1975) and Nadarajah and Zografos (2003, 2005).

In a similar way to the Csiszár generalization of φ-divergences we have the φ-entropies introduced by Burbea and Rao (1982a, 1982b, 1982c) and defined by

$$I_\varphi(X) \equiv I_\varphi(f) = \int \varphi(f) d\mu,$$

where φ is a continuous concave function defined on $(0, \infty)$, with $\varphi(0) = \lim_{u \searrow 0} \varphi(u) \in (-\infty, \infty]$. Some examples of the family of φ-entropies are provided in Table 1.3. In

Table 1.3: (φ)-Entropies

Entropy	$\varphi(x)$
Shannon (1948)	$-x\log x$
Havrda and Charvat (1967)	$(1-a)^{-1}(x^a-x),\ a\neq 1,\ a>0$
Kapur (1972)	$\frac{x^s+(1-x)^s-1}{1-s},\ s\neq 1$
Burbea (1984)	$\frac{x^s-(1+x)^s+1+(s-1)^{-1}(2^s-2)x}{s-2},\ s\neq 2$

order to include in the general family some additional measures, Salicrú et *al.* (1993) defined the (h,φ) entropy as

$$I_{h,\varphi}(X)\equiv I_{h,\varphi}(f)=h\left(\int\varphi(f)\,d\mu\right),$$

where $\varphi:(0,\infty)\to\mathbb{R}$ concave and $h:\mathbb{R}\to\mathbb{R}$ differentiable and increasing, or $\varphi:(0,\infty)\to\mathbb{R}$ convex and $h:\mathbb{R}\to\mathbb{R}$ differentiable and decreasing. Members of this family are given in Table 1.4.

Based on the φ entropy Burbea and Rao (1982a, 1982b, 1982c) defined the family of the R_φ-divergence

$$R_\varphi(f_1,f_2)=I_\varphi\left(\frac{f_1+f_2}{2}\right)-\frac{I_\varphi(f_1)+I_\varphi(f_2)}{2}$$

which was generalized by Pardo, L. et *al.* (1993) using the (h,φ) entropy to define the R_φ^h-divergence

$$R_\varphi^h(f_1,f_2)=I_\varphi^h\left(\frac{f_1+f_2}{2}\right)-\frac{I_\varphi^h(f_1)+I_\varphi^h(f_2)}{2}.$$

R_φ-divergence leads to another important family of divergences, the R_φ^α-divergence (Havrda and Charvat, 1967) which is obtained by φ-entropy using

$$\varphi(x)\equiv\varphi_\alpha(x)=\begin{cases}(1-\alpha)^{-1}(x^\alpha-x), & \alpha\neq 1,\ \alpha>0\\ -x\log x, & \alpha=1\end{cases}.$$

1.3 Fisher - Type Measures

Let X be a random variable with probability density function $f_\theta(x)$, that depends on a parameter θ (or a vector parameter θ) and corresponding distribution function

Table 1.4: (h, φ)-Entropies

Entropy	$\varphi(x)$	$h(x)$
Rényi (1961), $r \neq 0, 1$	x^r	$(r\,(1-r))^{-1} \log x$
Varma (1966), $0 < r < m,\ \ m \geqslant 1$	$x^{r/m}$	$(m\,(m-r))^{-1} \log x$
Arimoto (1971), $t \neq 1,\ \ t > 0$	$x^{1/t}$	$(t-1)^{-1}\,(x^t - 1)$
Sharma and Mittal (1977), $s \neq 1,\ \ s > 0$	$xlnx$	$\frac{\exp[(s-1)x]-1}{1-s}$
Sharma and Mittal (1977), $r \neq 1, s \neq 1, r > 0, s > 0$	x^r	$\frac{1}{1-s}\left(x^{\frac{s-1}{r-1}} - 1\right)$
Ferreri (1980), $\lambda > 0$	$(1+\lambda x)\log(1+\lambda x)$	$\left(1+\frac{1}{\lambda}\right)\log(1+\lambda) - \frac{x}{\lambda}$

P_θ. Let the parametric space Θ be an open subset of $\Re^k$, $k \geq 1$. Fisher information measure (Fisher, 1925)

$$FIS_X(\theta) = \int \left(\frac{\partial \log f_\theta(x)}{\partial \theta}\right)^2 f_\theta(x) d\mu, \quad f_\theta(x) = P_\theta / d\mu$$

is the representative and the most well known measure of this category. It measures *"the ease with which a parameter can be estimated"* (Lehmann, 1983), or *"the extent to which uncertainty is reduced by the observation"* (Rao, 1973). While Fisher's measure of information can be computed for any parametric family of distributions, it posseses interesting information theoretic and statistical properties provided that certain regularity conditions on $f_\theta(x)$ are satisfied (see e.g. Ferentinos and Papaioannou, 1981; Papaioannou, 1985).

Fisher information measure is connected to Kullback-Leibler divergence in the following setting. Let

$$f_1 = f_\theta \ and \ f_2 = f_{\theta+\Delta\theta},$$

where θ, $\Delta\theta$ are univariate neighboring points in the parametric space. Then,

$$I_X^{KL}(f_\theta, f_{\theta+\Delta\theta}) = 2\,(\Delta\theta)^2\,FIS_X(\theta)\,.$$

So $FIS_X(\theta)$ can be seen as a discrimination between neighboring points in the parametric space Θ.

If

$$\frac{d}{\partial\theta}E_\theta\Big(\frac{\partial}{\partial\theta}\log f_\theta(X)\Big)=\int\frac{\partial}{\partial\theta}\Big[\frac{\partial}{\partial\theta}(\log f_\theta(x))f_\theta(x)\Big]d\mu$$

then the Fisher information measure (see Casella and Berger, 2001, p. 338) takes the form

$$FIS_X(\theta)=-\int\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\,f_\theta(x)d\mu.$$

According to the above expression, information may be seen to be a measure of the "sharpness" of the support curve near the maximum likelihood estimate of θ.

A famous result involving Fisher information is the well known Cramér-Rao inequality, which states that the inverse of the Fisher information is an asymptotic lower bound of the variance of any unbiased estimator of θ. Another important result is that if $T=t(X)$ is a statistic, then

$$FIS_T(\theta)\leqslant FIS_X(\theta)$$

with equality if and only if T is a sufficient statistic.

A second form of Fisher's, is the so called Shift - Invariant Fisher Information presented as

$$I_X^F=E\left(\frac{\partial}{\partial x}\log f_\theta(x)\right)^2$$

or in a different form as

$$J_X^F=-E\left(\frac{\partial^2}{\partial x^2}\log f_\theta(x)\right).$$

This quantity was initially used by Rao (1958) for the determination of a lower bound analogous to Cramér-Rao. Recently Kagan (2001) had a similar approach for the Poisson distribution. Applications of this quantity in measuring the stochastic dependence of two or more random variables have been discussed by Zografos (1998, 2000). In physics and more specifically in optics and mechanics, Frieden (1988, 1998) refers to this quantity using the term "Extreme Physical Information".

Shift - Invariant Fisher Information I_X^F, also called Fisher information number, is not a measure of information, but it is a characteristic of a distribution and has other

Table 1.5: Fisher Information for typical Distributions

Distribution	Information measure	Information number
Normal, σ^2 known	$1/\sigma^2$	$1/\sigma^2$
Normal, μ known	$1/2\sigma^4$	$1/\sigma^2$
Normal, (μ, σ^2) known	$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}$	$1/\sigma^2$
Exponential (λ)	$1/\lambda^2$	λ^2

interesting properties. Note also that

$$I_X^F \neq J_X^F$$

and in fact

$$I_X^F + f_\theta'(a) - f_\theta'(b) = J_X^F,$$

where $f_\theta(x)$ is a probability density function and $a \leqslant x \leqslant b$.

Shift - Invariant Fisher Information coincides with the Fisher's information for a location parameter, namely

$$I_X^F = I_Y^F(\theta), \text{ if } Y = X + \theta.$$

Some examples are presented in Table 1.5 which has been reproduced from Papaioannou and Ferentinos (2005). For more about Shift - Invariant Fisher Information and its properties refer to Papaioannou and Ferentinos (2005).

1.4 Bayesian- Type Measures

The main representative of this type of measures is Lindley's Information Measure (1956) which will be presented in this section.

Consider the decision problem of reporting a distribution regarding an unknown parameter θ, belonging to a parametric space Θ, through an experiment $\boldsymbol{\varepsilon}$ that will result an observation x. In other words, we have a random variable X with probability density function $f(x)$ and an unknown quantity θ that we suppose it follows a prior distribution with density function f_θ. According to this notation we have

$$f(x) = \int\limits_{\Theta} f(x\,|\theta\,) f_\theta d\theta,$$

and Bayes' theorem reads

$$f(\theta\,|x\,) = \frac{f(x\,|\theta\,)\, f_\theta}{f(x)}.$$

The amount of information, before the experiment is performed, is defined to be

$$I_0 \equiv \int f_\theta \log f_\theta d\theta \equiv E_\theta \left[\log f_\theta\right].$$

After the completion of the experiment, the posterior distribution of θ is $f(\theta|x)$ and the amount of information becomes

$$I_1(x) \equiv \int f(\theta\,|x\,) \log f(\theta\,|x\,)\, d\theta.$$

Lindley's Information Measure is defined to be the average amount of information provided by an experiment $\boldsymbol{\varepsilon}$ with prior knowledge f_θ, as follows

$$\begin{aligned} I^L(\boldsymbol{\varepsilon}, f_\theta) &\equiv E_X\left[I_1(X) - I_0\right] \\ &= E_X E_\theta \left[\log\left\{\frac{f(\theta\,|X\,)}{f_\theta}\right\}\right] \\ &= E_X E_\theta \left[\log\left\{\frac{f(X\,|\theta\,)}{f(X)}\right\}\right] \\ &= \iint f(x,\theta) \log\left\{\frac{f(x,\theta)}{f(x)\, f_\theta}\right\} dx d\theta, \end{aligned}$$

where $f(x,\theta)$ is the joint density for X and θ.

For more details on Bayes risk based measures of the information in an experiment, see Lindley (1961), Chaloner and Verdinelli (1995), Dawid (1998), or Dawid and Sebastiani (1999).

1.5 The BHHJ Measure of Divergence

One of the most recently proposed measures of divergence is the BHHJ power divergence between f and g (Basu et *al.*, 1998) which is denoted by BHHJ, indexed by a positive parameter a, and defined as:

$$I_X^a(g,f) = \int \left\{ f^{1+a}(z) - \left(1+\frac{1}{a}\right) g(z)\, f^a(z) + \frac{1}{a} g^{1+a}(z) \right\} dz, \ \ a > 0. \quad (1.5.1)$$

Note that the above family which is also referred to as a family of power divergences is loosely related to the Cressie and Read power divergence (Basu et *al.*, 1998). This family of measures was proposed by Basu et *al.* (1998) for the development of a minimum divergence estimating method for robust parameter estimation. The index a controls the trade-off between robustness and asymptotic efficiency of the parameter estimators which are the quantities that minimize (1.5.1). It should be also noted that the BHHJ family reduces to the Kullback-Leibler divergence for a tending to 0 (see Lemma 2.2.3) and as it can be easily seen, to the square of the standard L_2 distance between f and g for $a = 1$. As a result, for $a = 0$ the family, as an estimating method, reduces to the traditional maximum likelihood estimation while for $a = 1$ becomes the mean squared error estimation. In the former case the resulting estimator is efficient but not robust while in the latter the method results in a robust but inefficient estimator. The authors observed that for values of a close to 0 the resulting estimators have strong robust features without a big loss in efficiency relative to the maximum likelihood estimating method. As a result one is interested in small values of $a \geq 0$, say between zero and one, although values larger than one are also allowed. One should be aware though of the fact that the estimating method becomes less and less efficient as the index a increases.

It is interesting to note that the BHHJ measure can be considered as a special case of the Bregman divergence (Jones and Byrne, 1990; Csiszár, 1991) which has the general form

$$\int \Big[H\{g(z)\} - H\{f(z)\} - \{g(z) - f(z)\}H'\{f(z)\}\Big]dz,$$

where H is a convex function. Observe that a Taylor series expansion of the integrand of the Bregman divergence when f is close to g gives

$$\frac{1}{2}(f-g)^2 H''(f).$$

If ones wants the Bregman divergence to reduce to the square of the L_2 distance for $a = 1$ (and consequently to the mean squared error estimating method) then $H''(f) \propto f^{a-1}$ for some $a \geq 0$ so that $H(f) \propto f^{a+1}$ in which case the Bregman divergence reduces to (1.5.1).

Some motivation for the form of the BHHJ divergence can be obtained by looking at the location model with location parameter θ. Note that in this case

$$\int f_\theta^{1+a}(z)dz$$

is independent of θ and the minimum divergence estimator is now the maximizer of

$$\sum_{i=1}^{n} f_\theta^a(X_i),$$

with the corresponding estimating equations having the form

$$\sum_{i=1}^{n} u_\theta\left(X_i\right) f_\theta^a\left(X_i\right) = 0, \tag{1.5.2}$$

where $u_\theta\left(z\right) = \partial \log f_\theta\left(z\right) / \partial\theta$ is the maximum likelihood score function. In the fully efficient case where $a = 0$, the estimating equation becomes $\sum\limits_{i=1}^{n} u_\theta\left(X_i\right) = 0$. For a random variable X in the exponential family with θ being the mean, $u_\theta(z) = (z-\theta)/\sigma^2$ where σ^2 the variance of X. Thus the sample mean is the MLE for θ, suggesting the robustness problems of maximum likelihood since all observations, including very severe outliers, get weights equal to one. On the other hand, when $a > 0$, and for several parametric models such as the normal, $u_\theta(z)f_\theta(z)$ is *bounded* function of z for fixed θ. As a result, (1.5.2) can be viewed as a weighted version of the efficient maximum likelihood score equation since it provides a relative-to-the-model downweighting for outlying observations; observations that are wildly discrepant with respect to the model will get nearly zero weights.

There can be no universal way of selecting an appropriate parameter a when applying the above estimating method. The value of a specifies the underlying distance measure and typically dictates to what extent the resulting method becomes statistically more robust than the maximum likelihood method, and should be thought of as an algorithmic parameter. A way of selecting the parameter a is by fixing the efficiency loss, at an ideal parametric model employed, at some low level, say 5%. Other ways could in some practical applications involve prior motions of the extent of contamination of the underlying model.

We generalize now the family (1.5.1) to a more general family of the following form that involves a general function $\Phi(\cdot)$.

Definition 1.5.1. *For a general function $\Phi \in \mathcal{G}$ and for $a > 0$ we define the divergence between two functions f and g by*

$$I_X^a\left(g,f\right) = E_g\Big(g^a(X)\Phi\Big(\frac{f(X)}{g(X)}\Big)\Big) \\ = \int g^{1+a}\left(z\right)\Phi\Big(\frac{f(z)}{g(z)}\Big)d\mu, \tag{1.5.3}$$

where μ represents the Lebesgue measure and $\mathcal{G}$ is the class of all convex functions Φ on $[0,\infty)$ such that $\Phi(1) = 0$, $\Phi'(1) = 0$ and $\Phi''\left(1\right) \neq 0$. In the expression of $I_X^a\left(g,f\right)$ we shall assume the conventions

$$0\Phi\left(0/0\right) = 0 \quad \text{and} \quad 0\Phi\left(u/0\right) = \lim_{u\to\infty}\Phi\left(u\right)/u, \text{ for } u > 0.$$

The BHHJ measure of Basu et. al (1998) can be obtained from the above general BHHJ family if the function Φ takes the special form

$$\Phi\left(u\right) = u^{1+a} - \left(1 + \frac{1}{a}\right)u^a + \frac{1}{a}\cdot \tag{1.5.4}$$

Expression (1.5.3) covers not only the continuous case presented in (1.5.1) but also a discrete setting where the measure μ is a counting measure. Indeed, for the discrete case, the divergence in (1.5.3) is meaningful for probability mass functions f and g whose support is a subset of the support S_μ, finite or countable, of the counting measure μ that satisfies

$$\mu(x) = 1 \; for \; x \in S_\mu$$

and 0 otherwise.

Consider now two discrete distributions $P = (p_1,\ldots,p_m)$ and $Q = (q_1,\ldots,q_m)$ with sample space $\Omega = \{x : p(x)\cdot q(x) > 0\}$, where $p(x),\ q(x)$ are the probability mass functions of the two distributions. Then the discrete version of the Cressie and Read measure is given by

$$I_X^{CR}\left(P,Q\right) = \frac{1}{\lambda\left(\lambda+1\right)}\sum_{i=1}^{m} p_i\Big[\Big(\frac{p_i}{q_i}\Big)^\lambda - 1\Big],\ \lambda \in R,\ \lambda \neq 0,-1. \tag{1.5.5}$$

The above measure was introduced by Cressie and Read (1984) for goodness of fit tests for multinomial distributions. Observe that the family includes important and well known test statistics like the Pearson's X^2 statistic (for $\lambda = 1$), the loglikelihood ratio statistic (for $\lambda \to 0$) and the Freeman-Tukey statistic (for $\lambda = -1/2$). Cressie and Read (1984) devoted their work to the analytic study of the asymptotic properties of the above measure and found that the $\lambda = 2/3$ case constitutes an excellent and compromising alternative between the traditional $\lambda \to 0$ (loglikelihood ratio test) and $\lambda = 1$ (Pearson's X^2 test) cases.

The discrete version of Csiszár's measure is given in a similar fashion, by

$$d_c = \sum_{i=1}^{m} q_i \varphi \left(p_i / q_i \right).$$

The discrete Csiszár's measure has been used by Zografos et *al.* (1990) for purposes analogous to the ones of the discrete Cressie and Read measure, namely for goodness of fit tests for multinomial distributions.

In what follows we extend the class of measures of divergence (1.5.3) to a discrete setting analogous to the above discrete versions of Csiszár's or Cressie and Read's measures for multinomial distributions.

Definition 1.5.2. *For two discrete distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ with sample space $\Omega = \{x : p(x) \cdot q(x) > 0\}$, where $p(x),\ q(x)$ are the probability mass functions of the two distributions, the discrete version of the general BHHJ family of divergence measures with a general function Φ as in Definition 1.5.1 and $a > 0$ is given by*

$$\begin{aligned} d_a \equiv d_a\left(Q, P\right) &= E_q\Big(q^a(X)\Phi\Big(\frac{p(X)}{q(X)}\Big)\Big) \\ &\equiv \sum_{i=1}^{m} q_i^{1+a} \Phi\left(\frac{p_i}{q_i}\right) \end{aligned} \tag{1.5.6}$$

which for Φ as in (1.5.4) becomes the discrete BHHJ measure given by

$$d_a \equiv d_a\left(Q, P\right) = \sum_{i=1}^{m} p_i^{1+a} - \left(1 + \frac{1}{a}\right) \sum_{i=1}^{m} q_i p_i^a + \frac{1}{a} \sum_{i=1}^{m} q_i^{1+a}. \tag{1.5.7}$$

Lemma 2.2.3 shows that for $a \to 0$ the measure reduces to the Kullback-Leibler divergence while for $\Phi(u) = \varphi(u)$ and for $a = 0$ we obtain the Csiszár's φ divergence.

The measures described in this chapter play a significant role in statistical inference and have several applications. In the rest of this thesis we focus on the general BHHJ family of divergence measures presented in Definitions 1.5.1 and 1.5.2 and investigate on one hand its basic properties and on the other hand its implementation in statistical modelling and in testing statistical hypotheses. For a review on measures of information see Papaioannou (2001). For a comprehensive discussion about statistical inference based on measures of divergence the reader is referred to Pardo (2006).

Chapter 2

Properties of the General BHHJ Family of Measures

2.1 Introduction

The measures of divergence are not formal distance functions. It is well known that any distance function $I(u, v)$ must satisfy the following three properties:

(1) $I(u, v) \geq 0$ with equality if and only if $u = v$

(2) $I(u, v) = I(v, u)$ and

(3) $I(u, w) \leq I(u, v) + I(v, w)$.

On the other hand any bivariate function $I(\cdot, \cdot)$ that satisfies the non-negativity property, namely $I(\cdot, \cdot) \geq 0$ with equality if and only if its two arguments are equal can possibly be used as a measure of information or divergence. Note that the Hellinger distance (the square root of the Matusita measure) given by

$$I_X^H(f, g) = (I_X^M(f, g))^{1/2} = \Big(\int (\sqrt{f} - \sqrt{g})^2 d\mu \Big)^{1/2}$$

is a true distance measure since it satisfies all three postulates.

Several properties of measures have been investigated over the years some of which are of axiomatic character and others of operational. By operational character we mean that the measures are involved in significant results in statistical inference (like the Cramér-Rao bound).

In this chapter we explore some of the basic properties of the general BHHJ family of measures of divergence with special attention given to the case where Φ is given by (1.5.4). In particular we discuss

- the nonnegativity property,
- the continuity property,
- the invariance property,
- the symmetry property,
- the limiting property,
- the order preserving property and
- the quadratic convergence.

For details about the properties mentioned in this chapter as well as about other properties of measures and information see Mathai and Rathie (1975), Ferentinos and Papaioannou (1981) and Papaioannou (1985).

2.2 Basic Properties

Let us define by $h(a)$ the integrand of $I_X^a(g,f)$ given in (1.5.1):

$$h(a) = f^{1+a}(z) - \left(1 + \frac{1}{a}\right) g(z) f^a(z) + \frac{1}{a} g^{1+a}(z).$$

The graphical representation of $h(a)$ is given in Figure 2.1. Observe that $h(-1) = 0$ and the maximum of $h(a)$ occurs for $a < 0$. Furthermore note that $h(0) = \lim_{a \downarrow 0} h(a) > h(a)$ for $a > 0$. Finally observe that

$$\exists\ a < 0,\ say\ a^*\ s.t.\ h(0) = h(a^*)\ and\ h(0) > h(a)\ for\ a < a^*.$$

Note that some of these characteristics are not valid for all functions f and g.

It is important to point out that for $a = 1$, the function $h(a)$ takes the form

$$h(1) = f^2 - 2gf + g^2$$

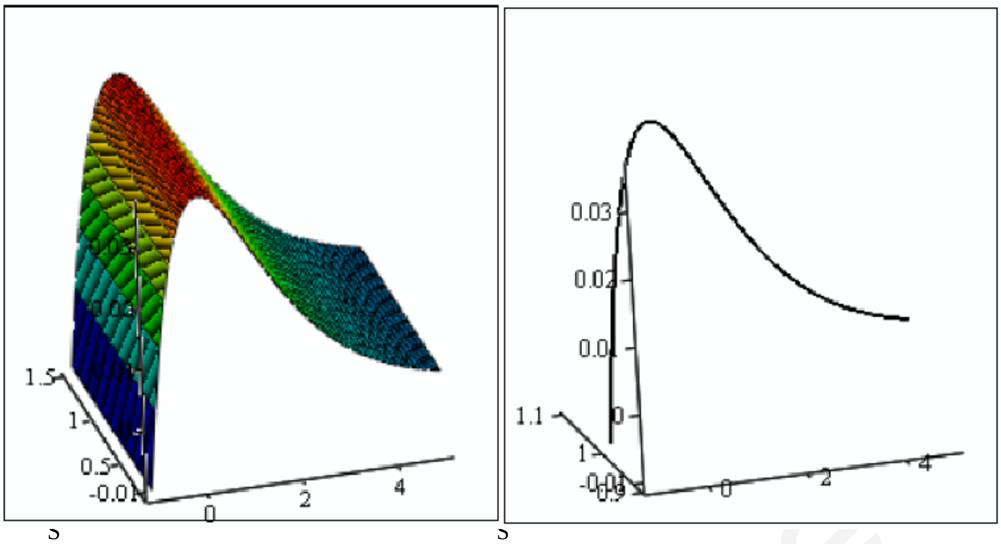

Figure 2.1: Graphical representation of $h(a)$ as a function of a and x (left graph) and as a function of a at $x = 1$ (right graph), where the distributions involved are Uniform(0,2) and Uniform(0,3).

so that the corresponding measure becomes the square of the standard L_2 distance, namely

$$I_X^1(g, f) = \int (f(z) - g(z))^2 dz.$$

Furthermore, although $h(a)$ is well defined for $a = -1$ this value is unacceptable due to the fact that the corresponding BHHJ measure between *any* functions f and g for $a = -1$ is meaningless.

It is easy to see that the BHHJ measure satisfies the basic properties of measures, namely the properties of nonnegativity and the continuity. In particular, as it was mentioned above, the value of measure is nonnegative for $a > -1$ while small changes in the distributions result in small changes in the measure. In other words, $I_X^a(\cdot, \cdot)$ is a continuous function in each of its arguments.

Finally, the value of the discrete measure is not affected by the simultaneous and equivalent reordering of the discrete masses in both the p_i's and the q_i's which confirms the invariance property of the discrete form of the BHHJ measure. Indeed, let $P_j = (p_{j_1}, \ldots, p_{j_m})$ and $Q_j = (q_{j_1}, \ldots, q_{j_m})$ reorderings of the original orderings of P and Q where $j = (j_1, \ldots, j_m)$ is an arbitrary permutation of the natural ordering of the set $(1, 2, \ldots, m)$. Then,

$$d_a(P_j, Q_j) = d_a(P, Q),$$

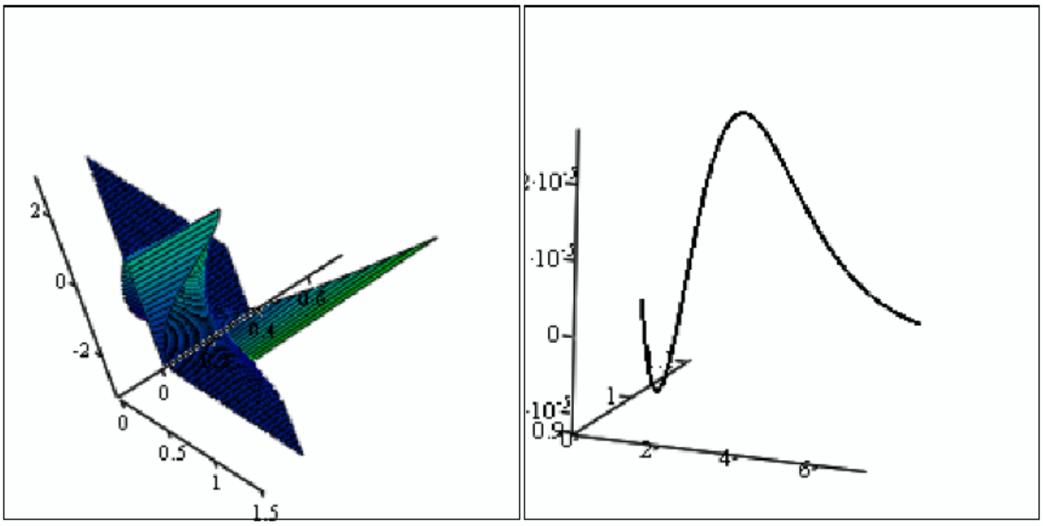

Figure 2.2: Graphical representation of function (2.2.1) as a function of a and x (left graph) and as a function of a at $x = 1$ (right graph), where the distributions involved are the Exponential with mean 2 and the Standard Normal.

for any reordering j.

For the symmetric property which is defined as

$$I_X^a(f, g) = I_X^a(g, f) \ or \ d_a(P, Q) = d_a(Q, P),$$

the following Lemma holds.

Lemma 2.2.1. *The symmetry property holds for the BHHJ measure for those values of a for which*

$$(1 - a)[g^{1+a}(x) - f^{1+a}(x)] + (1 + a)[f(x)g^a(x) - g(x)f^a(x)] = 0 \qquad (2.2.1)$$

$$(continuous \ case)$$

and

$$(1 - a)[q_i^{1+a} - p_i^{1+a}] + (1 + a)[p_i q_i^a - q_i p_i^a] = 0, \ \forall \ i \qquad (2.2.2)$$

$$(discrete \ case)$$

Proof. Using the definition of $I_X^a(f, g)$ and $I_X^a(g, f)$ it is easy to see that $I_X^a(f, g) = I_X^a(g, f)$ i.e.

$$\int \left\{ g^{1+a}(z) - \left(1 + \frac{1}{a}\right) f(z) g^a(z) + \frac{1}{a} f^{1+a}(z) \right\} dz =$$

$$= \int \left\{ f^{1+a}(z) - \left(1 + \frac{1}{a}\right) g(z) f^a(z) + \frac{1}{a} g^{1+a}(z) \right\} dz \qquad (2.2.3)$$

if the first of the above conditions is satisfied.

The discrete part is shown similarly for $d_a(P,Q)$ and $d_a(Q,P)$. ■

The graphical representation of the above function for the continuous case, as a function of α and x, appears in Figure (2.2) where the distributions involved are the Exponential with mean 2 and the Standard Normal. The figure implies that the Lemma holds true for $a=0$ and $a=1$ but only the second solution which is associated with the L_2 distance, is acceptable. The solution $a=0$ is unacceptable since (2.2.3) is not defined for $a=0$. Indeed, Lemma 2.2.3 shows that the BHHJ measure is defined for $a=0$ by continuity and in fact it reduces to the Kullback-Leibler measure which does not satisfy the symmetry property.

In Lemma 2.2.2 we investigate the limiting property according to which a sequence of probability density functions f_n converges to a probability density function f iff the corresponding measure of divergence $I_X^a(f_n,f)$ tends to 0. Before the statement of the Lemma we provide the definition of the μ-almost everywhere convergence, f_n to f:

Definition 2.2.1. *μ-almost everywhere convergence is a weakened version of pointwise convergence which states that, for X a measure space, $f_n(x) \to f(x)$ for all $x \in Y$, where Y is a measurable subset of X such that $\mu(X \backslash Y) = 0$.*

Lemma 2.2.2. *Let μ be a measure, Φ a function, f_n and f two probability density functions (pdfs) and $a > 0$ such that the following conditions hold*

- *I.* $\int \left| f^{1+a} \Phi\left(\frac{f_n}{f}\right) \right| d\mu < \infty$,
- *II.* Φ *is a continuous function,*
- *III.* $\Phi(1) = 0$, $\Phi'(1) = 0$, *and* Φ *is strictly convex,*
- *IV.* $f > 0$ *μ-almost everywhere.*

Then, the BHHJ family of measures satisfies the limiting property defined by

$$f_n \to f \ \ \mu - almost\ everywhere, \text{ iff } I_X^a(f_n,f) \to 0,$$

where f_n is a sequence of probability density functions, f is the limiting probability density function and $I_X^a(f_n,f)$ is the general BHHJ measure based on the two pdfs.

Proof. By (1.5.3) we have that

$$I_X^a\left(f,f_n\right)=\int f^{1+a}\Phi\left(\frac{f_n}{f}\right)d\mu,$$

with $\Phi(u)$ given in (1.5.4). Observe that if $f_n\to f$ $\mu - almost\ everywhere$ then

$$\begin{aligned}\lim_{n\to\infty} I_X^a\left(f_n,f\right)&=\lim_{n\to\infty}\int f^{1+a}\Phi\left(\frac{f_n}{f}\right)d\mu\\ &\overset{condition\ I}{=}\int f^{1+a}\lim_{n\to\infty}\Phi\left(\frac{f_n}{f}\right)d\mu\\ &\overset{condition\ II}{=}\int f^{1+a}\Phi\left(\lim_{n\to\infty}\frac{f_n}{f}\right)d\mu\\ &=\int f^{1+a}\Phi\left(1\right)d\mu\\ &=0.\end{aligned}$$

On the other hand, let

$$I_X^a\left(f_n,f\right)\to 0.$$

Then,

$$\lim_{n\to\infty}\int f^{1+a}\Phi\left(\frac{f_n}{f}\right)d\mu=0. \tag{2.2.4}$$

By condition III we have

$$\Phi(z)\geq 0. \tag{2.2.5}$$

By (2.2.4), (2.2.5) and condition IV we have

$$\lim_{n\to\infty}\Phi\left(\frac{f_n}{f}\right)=0,\ \mu - almost\ everywhere \tag{2.2.6}$$

and finally, by (2.2.6) and condition III we have

$$\lim_{n\to\infty}\frac{f_n}{f}=1,\ \mu - almost\ everywhere.$$

■

The following Lemma provides the relation between the BHHJ measure and the Kullback-Leibler measure. In particular, we show that for a tending to 0, the BHHJ measure reduces to the Kullback-Leibler measure.

Lemma 2.2.3. *The limit of (1.5.3) with $\Phi(u)$ as in (1.5.4) when $a\downarrow 0$ is the Kullback-Leibler divergence. Furthermore, the discrete form of the measure (1.5.6) tends to the discrete Kullback-Leibler measure given by*

$$d_a(P,Q)=\sum_{i=1}^{m}p_i\log\left(\frac{p_i}{q_i}\right)$$

for $a \downarrow 0$ and with $\Phi(u)$ as in (1.5.4).

Proof. The proof is given for (1.5.3). Observe that

$$\begin{aligned}
I_X^0\left(g,f\right) &= \lim_{a\downarrow 0} I_X^a\left(g,f\right)\\
&= \lim_{a\downarrow 0}\int\left\{f^{1+a}\left(z\right)-\left(1+\tfrac{1}{a}\right)g\left(z\right)f^a\left(z\right)+\tfrac{1}{a}g^{1+a}\left(z\right)\right\}dz\\
&= \lim_{a\downarrow 0}\int f^{1+a}\left(z\right)dz-\lim_{a\downarrow 0}\int g\left(z\right)f^a\left(z\right)dz+\lim_{a\downarrow 0}\int\tfrac{g(z)(g^a(z)-f^a(z))}{a}dz\\
&= \int f\left(z\right)dz-\int g\left(z\right)dz+\int g\left(z\right)\lim_{a\downarrow 0}\tfrac{(g^a(z)-f^a(z))}{a}dz\\
&= \int\Big(f\left(z\right)-g\left(z\right)\Big)dz+\int g\left(z\right)\lim_{a\downarrow 0}\left\{g^a\left(z\right)\log\left[g\left(z\right)\right]-f^a\left(z\right)\log\left[f\left(z\right)\right]\right\}dz\\
&= 1-1+\int g\left(z\right)\log\left\{\tfrac{g(z)}{f(z)}\right\}dz\\
&= \int g\left(z\right)\log\left\{\tfrac{g(z)}{f(z)}\right\}dz\\
&= I_X^{KL}(g,f).
\end{aligned}$$

■

We close this section with the order preserving property which has been introduced by Shiva, Ahmed and Georganas (1973) for entropy-type measures and states that the relation between the amount of information contained in a r.v X_1 and that contained in another r.v. X_2 remains intact irrespectively of the measure of information used.

The property was extended to Fisher-type measures by Papaioannou (1985) and to divergence measures by Zografos (1987). This property is natural in the sense that a measure of information of any type (entropy, information, divergence etc.) measures the amount of information available and therefore if a random variable contains a larger amount of information than another random variable for a specific measure then it is reasonable to expect that it will contain a larger amount of information for any measure.

In particular, if the superscripts (1) and (2) represent two different measures of information then

$$I_{X_1}^{(1)}(f_1,g_1)\geq I_{X_2}^{(1)}(f_2,g_2)\Leftrightarrow I_{X_1}^{(2)}(f_1,g_1)\geq I_{X_2}^{(2)}(f_2,g_2).$$

The following Lemma holds for the BHHJ divergence.

Lemma 2.2.4. *For some probability density functions f_1, f_2, g_1 and g_2, so that $I_X^\alpha(f_1,g_1)$ and $I_X^\alpha(f_2,g_2)$ are decreasing for $\alpha>0$, the following statements are equivalent:*

(a) $I_X^{\alpha}(f_1, g_1) \geq I_X^{\alpha}(f_2, g_2)$*, for* $\alpha \in (0, \alpha_3]$

(b) $I_X^{KL}(f_1, g_1) \geq I_X^{KL}(f_2, g_2)$

(c) $I_X^{R,\alpha}(f_1, g_1) \geq I_X^{R,\alpha}(f_2, g_2)$*, for* $\alpha \in (\alpha_1, \alpha_2)$

(d) $I_X^{Ka}(f_1, g_1) \geq I_X^{Ka}(f_2, g_2)$*, for* $\alpha_2 \geq 2$*,*

(e)$I_X^{M}(f_1, g_1) \geq I_X^{M}(f_2, g_2)$*, for* $\alpha_1 \leq 1/2$*,*

where α_1*,* α_2 *and* α_3 *are determined from the equations*

$$I_X^{R,\alpha_1}(f_1, g_1) = I_X^{KL}(f_2, g_2),$$

$$I_X^{R,\alpha_2}(f_2, g_2) = I_X^{KL}(f_1, g_1)$$

and

$$I_X^{KL}(f_2, g_2) = I_X^{\alpha_3}(f_1, g_1).$$

Proof. Part (b) follows immediately from part (a) if we take the limit as $a \to 0$ on both sides of

$$I_X^{\alpha}(f_1, g_1) \geq I_X^{\alpha}(f_2, g_2).$$

Then, the result follows from Lemma 2.2.3. Assume now that part (b) holds, namely

$$I_X^{KL}(f_1, g_1) \geq I_X^{KL}(f_2, g_2).$$

Since $I_X^{\alpha}(\cdot, \cdot)$ is a decreasing function of α, for $\alpha > 0$ and also

$$\lim_{\alpha \to 0} I_X^{\alpha}(\cdot, \cdot) = I_X^{KL}(\cdot, \cdot)$$

then

(i) $I_X^{KL}(f_2, g_2) \geq I_X^{\alpha}(f_2, g_2)$, for $\alpha > 0$

(ii) $\exists\ \alpha_3 > 0$ such that $I_X^{KL}(f_2, g_2) = I_X^{\alpha_3}(f_1, g_1)$ and

(iii) $I_X^{\alpha}(f_1, g_1) \geq I_X^{\alpha_3}(f_1, g_1)$ if $\alpha \in (0, \alpha_3]$.

Hence if $a \in (0, \alpha_3]$ and using (i) - (iii) we have

$$I_X^{\alpha}(f_1, g_1) \geq I_X^{\alpha_3}(f_1, g_1) = I_X^{KL}(f_2, g_2) \geq I_X^{\alpha}(f_2, g_2).$$

Parts (b)$-$(e) are equivalent from Theorem 2.1 and Corollary 2.1 of Zografos et *al.* (1989).

■

The above Lemma clearly shows that the key role in establishing the order preserving property is played by the parameter α involved in the measures examined. In particular, the property holds provided that the parameter α belongs to a specific interval, different for each measure. It should be noted that the end points of the interval depend on the distributions involved.

As a result, the order preserving property doesn't hold in a universal way for every measure and for every parameter α. Furthermore, the lemma by providing the range of values of the parameter α for which the measures describe properly (with consistency) the amount of information contained in the data, implies that the use of these measures should be limited to those values of α for which the order preserving property holds.

2.3 Quadratic Convergence of Discretized Versions of the BHHJ Measure

In practical situations the data are discrete or if they are continuous they are available in groups. In the latter case, the sample space is partitioned into disjoint intervals so that the theoretical distributions are approximated by the discrete distributions generated by these intervals. Several authors have considered this problem. Ghurye and Johnson (1981) showed that the discretized version of Kullback-Leibler divergence converges quadratically to the theoretical Kullback-Leibler measure. The same was proved by Zografos et *al.* (1986) for the Csiszár's φ-family of divergences as well as for the Rényi's and Fisher's measures. Both papers examined this discretization problem by considering a special partition of the sample space. In this section we generalize the above results by showing the quadratic convergence of the general BHHJ measure under suitable conditions and for the same special partition of the sample space.

Let us consider the following discretized partition of the sample space:

$$\Delta_{h,k} = \left(\left(k - \frac{1}{2}\right) h, \left(k + \frac{1}{2}\right) h\right), k = 0, \pm 1, \pm 2, ..., \quad h > 0.$$

Then the discretized versions of the functions f and g are given respectively by

$$p_k\left(h\right) = \int\limits_{\Delta_{h,k}} f\left(x\right)dx$$

and

$$q_k\left(h\right) = \int\limits_{\Delta_{h,k}} g\left(x\right)dx,\ k = 0, \pm1, \pm2, ...$$

Observe that the general BHHJ family of measures given in (1.5.3) can be written in the form

$$I_X^a(f,g) = \int\limits_{-\infty}^{+\infty} g^{1+a}(x)\Phi\left(\frac{f(x)}{g(x)}\right)dx$$

so that the discretized version becomes

$$I_h^a(f,g) = \sum_k q_k^{1+a}\left(h\right)\Phi\left(\frac{p_k\left(h\right)}{q_k\left(h\right)}\right).$$

The amount of information lost when using the discretized version $I_h^a(f,g)$ of $I_X^a(f,g)$ is given by:

$$D\left(h,a\right) = I_X^a(f,g) - I_h^a(f,g) = \sum_k J_k^a(h) \tag{2.3.1}$$

where

$$\begin{aligned} J_k^a(h) &= \int\limits_{\Delta_{h,k}} g^{1+a}(x)\Phi\left(\tfrac{f(x)}{g(x)}\right)dx - q_k^{1+a}(h)\Phi\left(\tfrac{p_k(h)}{q_k(h)}\right) \\ &= \int\limits_{\Delta_{h,k}} H_a(x)dx - q_k^{1+a}(h)\Phi\left(\tfrac{p_k(h)}{q_k(h)}\right) \end{aligned}$$

and

$$H_a(x) = g^{1+a}(x)\Phi\left(\frac{f(x)}{g(x)}\right).$$

Regularity Conditions for the quadratic convergence:

- I. $\int\limits_{-\infty}^{+\infty} g\left(x\right)\left|\Phi\left(f\left(x\right)/g\left(x\right)\right)\right|dx < \infty$

- II. f and g have the same support $S = \{x : f\left(x\right) > 0\} = \{x : g\left(x\right) > 0\}$, an open interval in $\Re$.

- III. f and g have continuous second derivatives on S; Φ has also a continuous second derivative on $(0,\infty)$.

- IV. The functions

$$g^{a}g''\Phi\left(\frac{f}{g}\right),\; g^{a-1}\left(g'\right)^{2}\Phi\left(\frac{f}{g}\right),\; g^{a}f''\Phi'\left(\frac{f}{g}\right),$$

$$g^{a-1}f'g'\Phi'\left(\frac{f}{g}\right),\; g^{a}f\left(\frac{g'}{g}\right)^{2}\Phi'\left(\frac{f}{g}\right),$$

$$fg^{a-1}g''\Phi'\left(\frac{f}{g}\right),\; \textit{and } g^{1+a}\left[\left(f/g\right)'\right]^{2}\Phi''\left(\frac{f}{g}\right)$$

are Riemann-integrable on $(-\infty,+\infty)$.

- V. If $x_{h,k}, y_{h,k}, z_{h,k} \in \Delta_{h,k} = \left(\left(k-1/2\right)h, \left(k+1/2\right)h\right), k = 0, \pm1, \pm2, ...$ and $a = a\left(h\right) = O\left(h^{\gamma}\right)$, with $\gamma < 1$ then

$$\lim_{h\to0} h\sum_{k} g^{a}\left(y_{h,k}\right)g''\left(x_{h,k}\right)\Phi\left(\frac{f\left(y_{h,k}\right)}{g\left(y_{h,k}\right)}\right) = \int\limits_{-\infty}^{+\infty} g''\left(x\right)\Phi\left(\frac{f\left(x\right)}{g\left(x\right)}\right)dx,$$

$$\lim_{h\to0} h\sum_{k} g^{a}\left(z_{h,k}\right)f''\left(x_{h,k}\right)\Phi'\left(\frac{f\left(y_{h,k}\right)}{g\left(y_{h,k}\right)}\right) = \int\limits_{-\infty}^{+\infty} f''\left(x\right)\Phi'\left(\frac{f\left(x\right)}{g\left(x\right)}\right)dx,$$

$$\lim_{h\to0} h\sum_{k} \frac{g''\left(x_{h,k}\right)f\left(y_{h,k}\right)}{g^{1-a}\left(y_{h,k}\right)}\Phi'\left(\frac{f\left(z_{h,k}\right)}{g\left(z_{h,k}\right)}\right) = \int\limits_{-\infty}^{+\infty} \frac{g''\left(x\right)f\left(x\right)}{g\left(x\right)}\Phi'\left(\frac{f\left(x\right)}{g\left(x\right)}\right)dx,$$

$$\lim_{h\to0} h\sum_{k} g^{1+a}\left(x_{h,k}\right)\left\{\left(\frac{f\left(x_{h,k}\right)}{g\left(x_{h,k}\right)}\right)'\right\}^{2}\Phi''\left(\frac{f\left(x_{h,k}\right)}{g\left(x_{h,k}\right)}\right) =$$
$$\int\limits_{-\infty}^{+\infty} g\left(x\right)\left\{\left(\frac{f(x)}{g(x)}\right)'\right\}^{2}\Phi''\left(\frac{f(x)}{g(x)}\right)dx.$$

Theorem 2.3.1. *Under the regularity conditions stated in the present section we have the following result*

$$
\begin{aligned}
&h^{-2}D\left(h,a\right)= \\
&=h^{-1}\left(1-h^{a}\right)\sum_{k}g^{1+a}(kh)\Phi\left(\tfrac{f(kh)}{g(kh)}\right) \\
&+\tfrac{h}{24}a\left(1+a\right)\sum_{k}g^{a-1}\left(u_{h,k}\right)\left(g'\left(u_{h,k}\right)\right)^{2}\Phi\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\
&+\left(1+a\right)\left[\tfrac{h}{24}\sum_{k}g^{a}\left(u_{h,k}\right)g''\left(u_{h,k}\right)\Phi\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)-\tfrac{h^{1+a}}{24}\sum_{k}g^{a}\left(kh\right)g''(m_{h,k})\Phi\left(\tfrac{f(kh)}{g(kh)}\right)\right] \\
&+\tfrac{h}{24}\sum_{k}g^{a}\left(u_{h,k}\right)f''\left(u_{h,k}\right)\Phi'\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)-\tfrac{h^{1+a}}{24}\sum_{k}g^{a}\left(kh\right)f''(v_{h,k})\Phi'\left(\tfrac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}\right) \\
&+\tfrac{h}{24}\sum_{k}2ag^{a-1}\left(u_{h,k}\right)f'\left(u_{h,k}\right)g'\left(u_{h,k}\right)\Phi'\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\
&-\tfrac{h}{24}\sum_{k}2ag^{a}\left(u_{h,k}\right)f\left(u_{h,k}\right)\left(\tfrac{g'\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)^{2}\Phi'\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\
&-\tfrac{h}{24}\sum_{k}f\left(u_{h,k}\right)g^{a-1}\left(u_{h,k}\right)g''\left(u_{h,k}\right)\Phi'\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\
&-\tfrac{h^{1+a}}{24}\sum_{k}f\left(kh\right)g^{a-1}\left(kh\right)g''(m_{h,k})\Phi'\left(\tfrac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}\right) \\
&+\tfrac{h}{24}\sum_{k}g^{1+a}\left(u_{h,k}\right)\left[\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)'\right]^{2}\Phi''\left(\tfrac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)+O\left(h^{3+a}\right)+O\left(h^{5+a}\right),
\end{aligned}
$$

where $u_{h,k},v_{h,k},m_{h,k},w_{h,k},\in\Delta_{h,k}$. *Furthermore, if* $a=a\left(h\right)=O\left(h^{\gamma}\right)$, *with* $\gamma<1$ *we have the quadratic convergence:*

$$\lim_{h\to0}\frac{I_{X}^{a}(f,g)-I_{h}^{a}(f,g)}{h^{2}}=\frac{1}{24}\int\limits_{-\infty}^{+\infty}g\left(x\right)\left[\left(\frac{f\left(x\right)}{g\left(x\right)}\right)'\right]^{2}\Phi''\left(\frac{f\left(x\right)}{g\left(x\right)}\right)dx.$$

Proof. Using Taylor theorem in symmetric form,

$$\omega\left(u+\frac{h}{2}\right)-\omega\left(u-\frac{h}{2}\right)=h\omega'(u)+\frac{h^{3}}{24}\omega'''(\bar{u})$$

where $\bar{u}\in\left(u-\frac{h}{2},u+\frac{h}{2}\right)$ and if ω''' exists and is continuous, we obtain

$$\int\limits_{\Delta_{h,k}}H_{a}(x)dx=hH_{a}(kh)+\frac{h^{3}}{24}H_{a}''\left(u_{h,k}\right),\ u_{h,k}\in\Delta_{h,k}\qquad(2.3.2)$$

where

$$
\begin{aligned}
H_{a}''=\ &\left(1+a\right)g^{a}\left(\tfrac{a(g')^{2}}{g}+g''\right)\Phi\left(\tfrac{f}{g}\right) \\
&+g^{1+a}\left[\left(\tfrac{f}{g}\right)'\right]^{2}\Phi''\left(\tfrac{f}{g}\right) \\
&+g^{a}\left(f''+2a\tfrac{f'.g'}{g}-2af\left(\tfrac{g'}{g}\right)^{2}-\tfrac{f.g''}{g}\right)\Phi'\left(\tfrac{f}{g}\right).
\end{aligned}
$$

A Taylor series expansion of $\Phi\left(\cdot\right)$ around the point $f\left(kh\right)/g\left(kh\right)$ yields

$$\Phi\left(\frac{p_{k}\left(h\right)}{q_{k}\left(h\right)}\right)=\Phi\left(\frac{f\left(kh\right)}{g\left(kh\right)}\right)+\left(\frac{p_{k}\left(h\right)}{q_{k}\left(h\right)}-\frac{f(kh)}{g(kh)}\right)\Phi'\left(r\right)\qquad(2.3.3)$$

with r a point belonging to the interval determined by the points $f\left(kh\right)/g\left(kh\right)$ and $p_k\left(h\right)/q_k\left(h\right)$. In fact, since f/g is continuous in all $\Delta_{h,k}$ partitions, $h>0$, $k=0,\pm 1,\pm 2,...$, belonging to S, we can easily see that there exists $w_{h,k}\in\Delta_{h,k}$ such that

$$r=\frac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}. \tag{2.3.4}$$

Also, for $v_{h,k},m_{h,k}\in\Delta_{h,k}$ we have

$$p_k\left(h\right)=hf\left(kh\right)+\frac{h^3}{24}f''\left(v_{h,k}\right)$$

and

$$q_k\left(h\right)=hg\left(kh\right)+\frac{h^3}{24}g''\left(m_{h,k}\right).$$

Using the binomial expansion for $x\ll 1$, namely,

$$\left(1+x\right)^p=1+px+O\left(x^2\right),\ \forall\ p,$$

we have

$$q_k^p\left(h\right)=h^pg^p\left(kh\right)+p\frac{h^{2+p}}{24}g''\left(m_{h,k}\right)g^{p-1}\left(kh\right)+O\left(h^{4+p}\right),\forall\ p. \tag{2.3.5}$$

Multiplying both sides of (2.3.3) by $q_k^p\left(h\right)$ with $p=1+a$ and using (2.3.4) and (2.3.5) we have

$$\begin{aligned} q_k^{1+a}\left(h\right)\Phi\left(\tfrac{p_k(h)}{q_k(h)}\right) = \; & h^{1+a}g^{1+a}\left(kh\right)\Phi\left(\tfrac{f(kh)}{g(kh)}\right)\\ & +\left(1+a\right)\tfrac{h^{3+a}}{24}g^a\left(kh\right)g''(m_{h,k})\Phi\left(\tfrac{f(kh)}{g(kh)}\right)\\ & +\tfrac{h^{3+a}}{24}g^a\left(kh\right)f''(v_{h,k})\Phi'\left(\tfrac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}\right)\\ & -\tfrac{h^{3+a}}{24}f\left(kh\right)g^{a-1}\left(kh\right)g''(m_{h,k})\Phi'\left(\tfrac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}\right)\\ & +O\left(h^{5+a}\right). \end{aligned}$$

Substituting the above formula along with (2.3.2) into (2.3.1) we have the form of

$D(h,a)$, namely

$$\begin{aligned} D\left(h,a\right) &= I_X^a(f,g) - I_h^a(f,g) = \sum_k J_k^a(h) \\ &= h\left(1-h^a\right)\sum_k g^{1+a}(kh).\Phi\left(\tfrac{f(kh)}{g(kh)}\right) \\ &+\tfrac{h^3}{24}a\left(1+a\right)\sum_k g^{a-1}\left(u_{h,k}\right)\left(g'\left(u_{h,k}\right)\right)^2\Phi\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\ &+\left(1+a\right)\left[\tfrac{h^3}{24}\sum_k g^a\left(u_{h,k}\right)g''\left(u_{h,k}\right)\Phi\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)\right. \\ &\left.-\tfrac{h^{3+a}}{24}\sum_k g^a\left(kh\right)g''(m_{h,k})\Phi\left(\tfrac{f(kh)}{g(kh)}\right)\right] \\ &+\tfrac{h^3}{24}\sum_k g^a\left(u_{h,k}\right)f''\left(u_{h,k}\right)\Phi'\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)-\tfrac{h^{3+a}}{24}\sum_k g^a\left(kh\right)f''(v_{h,k})\Phi'\left(\frac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}\right) \\ &+\tfrac{h^3}{24}\sum_k 2ag^{a-1}\left(u_{h,k}\right)f'\left(u_{h,k}\right).g'\left(u_{h,k}\right)\Phi'\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\ &-\tfrac{h^3}{24}\sum_k 2ag^a\left(u_{h,k}\right)f\left(u_{h,k}\right)\left(\frac{g'\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)^2\Phi'\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\ &-\tfrac{h^3}{24}\sum_k f\left(u_{h,k}\right)g^{a-1}\left(u_{h,k}\right).g''\left(u_{h,k}\right)\Phi'\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\ &-\tfrac{h^{3+a}}{24}\sum_k f\left(kh\right)g^{a-1}\left(kh\right)g''(m_{h,k})\Phi'\left(\frac{f\left(w_{h,k}\right)}{g\left(w_{h,k}\right)}\right) \\ &+\tfrac{h^3}{24}\sum_k g^{1+a}\left(u_{h,k}\right)\left[\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right)'\right]^2\Phi''\left(\frac{f\left(u_{h,k}\right)}{g\left(u_{h,k}\right)}\right) \\ &+O\left(h^{5+a}\right). \end{aligned}$$

Taking $a = a\left(h\right) = O\left(h^\gamma\right)$, with $\gamma < 1$, multiplying by h^{-2} and using Regularity Condition (V), we have the desired result, namely

$$\begin{aligned} \lim_{h\to 0}\tfrac{1}{h^2}D\left(h,a\right) &= \lim_{h\to 0}\tfrac{I_X^a(f,g)-I_h^a(f,g)}{h^2} \\ &= \lim_{h\to 0}\tfrac{1}{h^2}\sum_k J_k^a(h) \\ &= \tfrac{1}{24}\int_{-\infty}^{+\infty} g\left(x\right)\left[\left(\tfrac{f(x)}{g(x)}\right)'\right]^2\Phi''\left(\tfrac{f(x)}{g(x)}\right)dx. \qquad \blacksquare \end{aligned}$$

The regularity conditions of this theorem are generalizations of the conditions used by Ghurye and Johnson (1981) for the Kullback-Leibler measure and by Zografos et *al.* (1986) for the Csiszár's measure. As expected for various functions Φ we have different measures of divergence. More specifically for $\Phi = \varphi$ and $a = 0$ the above result reduces to the result for Csiszár's family of measures obtained by Zografos et *al.* (1986). Furthermore, for $a \to 0$, the result of Ghurye and Johnson (1981) is obtained. Notice that the same result can be obtained as a special case of Csiszár's measure for $\Phi(u) = u\log u$ and $a = 0$.

Other measures covered by the above theorem are the Kagan, the Matusita measure and the Vajda [Vajda, 1973] measure given by

$$I_X^V(f,g) = \int g(y)\left|1 - \frac{f(y)}{g(y)}\right|^{\beta} dy, \ \beta \geq 1.$$

Observe that the Vajda measure reduces to Kagan's measure for $\beta = 2$.

Chapter 3

Model Selection Criteria

3.1 Introduction

Since the measures of divergence are used as indices of similarity or dissimilarity between populations and for measuring mutual information concerning two variables they can be used for the construction of model selection criteria. A model selection criterion can be considered as an approximately unbiased estimator of the expected overall discrepancy, a nonnegative quantity which measures the *distance* between the true unknown model and a fitted approximating model. If the value of the criterion is small then the approximated model is good.

The Kullback-Leibler measure was the one used by Akaike (1973) to develop the Akaike Information Criterion (AIC). Let $\mathbf{x} = (x_1, \ldots, x_n)$ a realization of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ and assume that the X_i's are independent and identically distributed each with true unknown density function $g(\cdot, \theta_0)$, with $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})'$ the true but unknown value of the p-dimensional parameter of the distribution. Consider a candidate model $f_\theta(\cdot)$ and let $\hat{\theta}$ the maximum likelihood estimator (MLE) of θ_0 in some hypothesized set Θ, i.e.

$$l(\hat{\theta}; x) = \sum_{i=1}^{n} \log(f_{\hat{\theta}}(x_i)) = \max_{\theta \in \Theta} l(\theta; x)$$

so that $f_{\hat{\theta}}(\cdot)$ is an estimate of $g(\cdot, \theta_0)$. The divergence between the estimate (candidate model) and the true density can be measured by the Kullback-Leibler measure:

$$I_X^{KL}(g, f_{\hat{\theta}}) = \int g(y, \theta_0) \log\Big(\frac{g(y, \theta_0)}{f_{\hat{\theta}}(y)}\Big) dy$$

which is a special case for $a \to 0$ (see Lemma 2.2.3) of the BHHJ measure

$$I_X^a\left(g, f_{\hat{\theta}}\right) = \int \left\{ f_{\hat{\theta}}^{1+a}(z) - \left(1 + \frac{1}{a}\right) g(z) f_{\hat{\theta}}^{a}(z) + \frac{1}{a} g^{1+a}(z) \right\} dz. \qquad (3.1.1)$$

Observe that $I_X^{KL}(g, f_{\hat{\theta}})$ can be written in the form

$$I_X^{KL}(g, f_{\hat{\theta}}) = E_g[\log(g(X, \theta_0)] - E_g[\log(f_{\hat{\theta}}(X))].$$

Note that the first term is independent of the candidate model and therefore the divergence can be evaluated using only the second term, usually known as the expected loglikelihood. Akaike proposed the evaluation of the fit of $f_{\hat{\theta}}(\cdot)$ using minus twice the *mean* expected loglikelihood given by

$$-2E_g\left[E_g[\log(f_{\hat{\theta}}(X))]\right] = -2 \int \ldots \int E_g[\log(f_{\hat{\theta}}(X))] \prod_{i=1}^{n} g(x_i, \theta_0) dx_1 \ldots dx_n$$

since the candidate model is close to the true model if the above quantity is small. Furthermore, Akaike provided an unbiased estimator of the expected loglikelihood given by

$$[-2l(\hat{\theta}; x) + 2p]/n$$

so that the resulting AIC is defined to be

$$AIC = -2l(\hat{\theta}; x) + 2p.$$

A general class of criteria has been introduced by Konishi and Kitagawa (1996) which also estimates the Kullback-Leibler measure where the estimation is not necessarily based on maximum likelihood.

Following the early work of Akaike, other model selection proposals include Bayesian approaches with the Bayesian Information Criterion (BIC, Schwarz, 1978) and the Deviance Information Criterion (DIC, Spiegelhalter et *al.*, 2002; van der Linde, 2005) being the most popular. The BIC criterion has a number of advantages worth mentioning. More specifically, it has been shown to be consistent (Schwarz, 1978; Wei, 1982) which means that it chooses the correct model with probability 1 as n tends to infinity. The second advantage is that the criterion depends on $\log n$ instead of n and therefore it downweights the effective sample size which in some cases prevents the erroneous rejection of null hypothesis for large sample sizes.

Here we apply the same methodology used for AIC to the BHHJ divergence in order to develop a new criterion, the Divergence Information Criterion (DIC). Note that the DIC proposed here is not related to the above mentioned deviance information criterion which is a Bayesian criterion for posterior predictive comparisons.

3.2 The Construction of the New Criterion

Consider a random sample $X_1, \ldots, X_n$ from the distribution g (the true model) and a candidate model f_θ from a parametric family of models $\{f_\theta\}$, indexed by an unknown parameter $\theta \in \Theta$, where Θ is an one dimensional parametric space. To construct the new criterion for goodness of fit we shall consider the quantity:

$$W_\theta = \int \left\{ f_\theta^{1+a}(z) - \left(1 + a^{-1}\right) g(z)\, f_\theta^{a}(z) \right\} dz, \ a > 0 \tag{3.2.1}$$

which is the same as the BHHJ divergence $I_X^a(g, f_\theta)$ given in (1.5.1) without the last term that remains constant irrespectively of the model f_θ used. Observe that (3.2.1) can also be written as:

$$W_\theta = E_{f_\theta}\left(f_\theta^{a}(Z)\right) - \left(1 + a^{-1}\right) E_g\left(f_\theta^{a}(Z)\right), \ a > 0. \tag{3.2.2}$$

3.2.1 The Expected Overall Discrepancy

The target theoretical quantity that needs to be later approximated by an unbiased estimator is given by

$$EW_{\hat{\theta}} = E\left(W_\theta \middle| \theta = \hat{\theta}\right) \tag{3.2.3}$$

where $\hat{\theta}$ is any consistent and asymptotically normal estimator of θ. This quantity can be viewed as the average distance between g and f_θ up to a constant and is known as *the expected overall discrepancy between g and f_θ.*

Observe that the expected overall discrepancy can be easily evaluated by using a Taylor expansion around θ_0. The necessary derivatives of (3.2.2) are given below.

Lemma 3.2.1. *The first and second derivatives of (3.2.2) are:*

$$\frac{\partial W_\theta}{\partial \theta} = (a+1)\left[\int u_\theta(z)\, f_\theta^{1+a}(z)\, dz - E_g\left(u_\theta(Z)\, f_\theta^{a}(Z)\right)\right]$$

and

$$\frac{\partial^2 W_\theta}{\partial \theta^2} = (a+1)\left\{(a+1)\int \left[u_\theta\left(z\right)\right]^2 f_\theta^{1+a}\left(z\right)dz - \int i_\theta f_\theta^{1+a}\left(z\right)dz \right.$$
$$\left. +E_g\left(i_\theta\left(Z\right) f_\theta^{a}\left(Z\right)\right) - E_g\left(a\left[u_\theta\left(Z\right)\right]^2 f_\theta^{a}\left(Z\right)\right)\right\}$$

where $u_\theta\left(z\right) = \frac{\partial}{\partial\theta}\left(\log\left(f_\theta\left(z\right)\right)\right)$ *and* $i_\theta\left(z\right) = -\frac{\partial^2}{\partial\theta^2}\left(\log\left(f_\theta\left(z\right)\right)\right)$.

Proof. For the first derivative we have

$$\begin{aligned}\frac{\partial W_\theta}{\partial\theta} &= (a+1)\int f_\theta^{a}\left(z\right) f_\theta'\left(z\right)dz - \left(\frac{a+1}{a}\right) E_g\left(a f_\theta^{a-1}\left(Z\right) f_\theta'\left(Z\right)\right)\\ &= (a+1)\left[\int \frac{\partial}{\partial\theta}\left(\log f_\theta\left(z\right)\right) f_\theta^{1+a}\left(z\right)dz - E_g\left(\frac{\partial}{\partial\theta}\left(\log f_\theta\left(Z\right)\right) f_\theta^{a}\left(Z\right)\right)\right]\\ &= (a+1)\left[\int u_\theta\left(z\right) f_\theta^{1+a}\left(z\right)dz - E_g\left(u_\theta\left(Z\right) f_\theta^{a}\left(Z\right)\right)\right].\end{aligned}$$

Observe also that

$$\begin{aligned}\frac{\partial^2 W_\theta}{\partial\theta^2} &= (a+1)\left\{\int \left[-i_\theta\left(z\right) f_\theta^{1+a}\left(z\right) + u_\theta\left(z\right)\left(a+1\right) f_\theta^{a}\left(z\right) f_\theta'\left(z\right)\right]dz\right.\\ &\left. - E\ _g\left(-i_\theta\left(Z\right) f_\theta^{a}\left(Z\right) + a\left[u_\theta\left(Z\right)\right] f_\theta^{a-1}\left(Z\right) f_\theta'\left(Z\right)\right)\right\}\\ &= (a+1)\left\{\int \left[(a+1)\left[u_\theta\left(z\right)\right]^2 f_\theta^{1+a}\left(z\right) - i_\theta\left(z\right) f_\theta^{1+a}\left(z\right)\right]dz\right.\\ &\left. - E\ _g\left(-i_\theta\left(Z\right) f_\theta^{a}\left(Z\right) + a\left[u_\theta\left(Z\right)\right]^2 f_\theta^{a}\left(Z\right)\right)\right\}.\\ &= (a+1)\left\{(a+1)\int \left[u_\theta\left(z\right)\right]^2 f_\theta^{1+a}\left(z\right)dz - \int i_\theta f_\theta^{1+a}(z)dz\right.\\ &\left. + E_g\left(i_\theta\left(Z\right) f_\theta^{a}\left(Z\right)\right) - aE\ _g\left(\left[u_\theta\left(Z\right)\right]^2 f_\theta^{a}\left(Z\right)\right)\right\}.\end{aligned}$$

■

Lemma 3.2.2. *If the true distribution* g *belongs to the parametric family* $\{f_\theta\}$*, then the second derivative of (3.2.2) simplifies to:*

$$\frac{\partial^2 W_{\theta_0}}{\partial{\theta_0}^2} = (a+1)\int \left[u_{\theta_0}\left(z\right)\right]^2 f_{\theta_0}^{1+a}\left(z\right)dz = (a+1)\,J(\theta_0) \tag{3.2.4}$$

where $J(\theta_0) = \int \left[u_{\theta_0}\left(z\right)\right]^2 f_{\theta_0}^{1+a}\left(z\right)dz$ *and* θ_0 *represents the best fitting value of the parameter. Also the first derivative of (3.2.2), under the same assumption, is equal to 0.*

Proof.

If the true distribution g belongs to the parametric family $\{f_\theta\}$, then:

$$E_g\left(\left[u_{\theta_0}\left(Z\right)\right]^2 f_{\theta_0}^{a}\left(Z\right)\right) = \int \left[u_{\theta_0}\left(z\right)\right]^2 f_{\theta_0}^{1+a}\left(z\right)dz$$

and

$$E_g\left(i_{\theta_0}(Z)\, f_{\theta_0}^{a}(Z)\right) = \int i_{\theta_0}(z)\, f_{\theta_0}^{1+a}(z)dz$$

so that

$$\frac{\partial^2 W_{\theta_0}}{\partial{\theta_0}^2} = (a+1)\, J(\theta_0).$$

For the first derivative the result follows immediately since

$$E_g\left(u_{\theta_0}(Z) f_{\theta_0}^{a}(Z)\right) = \int u_{\theta_0}(z) f_{\theta_0}^{1+a}(z)\, dz \Rightarrow \frac{\partial W_{\theta_0}}{\partial \theta_0} = 0. \tag{3.2.5}$$

■

Theorem 3.2.1. *Under the assumptions of Lemma 3.2.1 the expected overall discrepancy at $\theta = \hat{\theta}$ is given by*

$$EW_{\hat{\theta}} = W_{\theta_0} + \frac{(a+1)}{2} E\left[\left(\hat{\theta} - \theta_0\right)^2 J(\theta_0)\right] + ER_n, \tag{3.2.6}$$

where $R_n = o((\hat{\theta} - \theta_0)^2)$, θ_0 the true value of the parameter and

$$J(\theta_0) = \int \left[u_{\theta_0}(z)\right]^2 f_{\theta_0}^{1+a}(z)\, dz.$$

Proof. Using a Taylor expansion of the quantity W_θ around the true parameter θ_0 and equation (3.2.4) and taking $\theta = \hat{\theta}$, W_θ simplifies to:

$$W_{\hat{\theta}} = W_{\theta_0} + \frac{(a+1)}{2}\left(\hat{\theta} - \theta_0\right)^2 J(\theta_0) + o((\hat{\theta} - \theta_0)^2). \tag{3.2.7}$$

It is easily seen that the expectation of $W_{\hat{\theta}}$ is given by (3.2.6). ■

The assumption that the true distribution g belongs to the parametric family $\{f_\theta\}$ is the assumption made by Akaike (Akaike, 1973). The assumption may be questionable in practice but it is a useful one in the sense that provides the basis for the evaluation of the estimator of the expected overall discrepancy as well as the computation of expectations for central distributions which would not have been possible otherwise (see also McQuarrie and Tsai, 1998, p. 20-21).

3.2.2 Estimation of the Expected Overall Discrepancy

In this section we construct an unbiased estimator of the expected overall discrepancy (3.2.6). First we shall deal with the estimation of the unknown density g. An estimator of (3.2.2) with respect to g is given by replacing $E_g\left(f_\theta^a(Z)\right)$ by its sample analogue

$$Q_\theta = \int f_\theta^{1+a}(z)\,dz - \left(1+\frac{1}{a}\right)\frac{1}{n}\sum_{i=1}^{n} f_\theta^a(X_i)\,. \tag{3.2.8}$$

The derivatives of Q_θ are given in the following lemma.

Lemma 3.2.3. *The derivatives of (3.2.8) are:*

$$\frac{\partial Q_\theta}{\partial\theta} = (a+1)\left[\int u_\theta(z)\,f_\theta^{1+a}(z)\,dz - \frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i)\,f_\theta^a(X_i)\right]$$

$$\text{and}\quad \frac{\partial^2 Q_\theta}{\partial\theta^2} = (a+1)\left\{(a+1)\int\left[u_\theta(z)\right]^2 f_\theta^{1+a}(z)\,dz - \right.$$

$$\left.\int i_\theta f_\theta^{1+a}(z)\,dz + \frac{1}{n}\sum_{i=1}^{n} i_\theta(X_i)\,f_\theta^a(X_i) - \frac{1}{n}\sum_{i=1}^{n} a\left[u_\theta(X_i)\right]^2 f_\theta^a(X_i)\right\},$$

where $u_\theta(z)$ and $i_\theta(z)$ are as in Lemma 3.2.1.

Proof. The proof is very similar to the proof of Lemma 3.2.1 and is omitted. ■

The Taylor expansion of the quantity Q_θ around the estimator $\hat\theta$ yields the approximation:

$$Q_\theta = Q_{\hat\theta} + \left(\theta-\hat\theta\right)\left[\frac{\partial Q_\theta}{\partial\theta}\right]_{\hat\theta} + \frac{1}{2}\left(\theta-\hat\theta\right)^2\left[\frac{\partial^2 Q_\theta}{\partial\theta^2}\right]_{\hat\theta} + o((\hat\theta-\theta)^2). \tag{3.2.9}$$

Recall that the estimator $\hat\theta$ is a consistent and asymptotically normal estimator of the parameter θ. For such an estimator one could select the value of θ that either maximizes the loglikelihood function (MLE method) or minimizes the BHHJ discrepancy or equivalently the quantity W_θ (Basu method). In the latter case the consistency as well as the asymptotic normality are verified by the theorem below which is due to Basu et *al.* (1998).

Theorem 3.2.2 (Basu et al. (1998)). *Under certain regularity conditions, there exists $\hat\theta$ such that, as $n\to\infty$,*

(i) $\hat\theta$ is consistent for θ_0, and

(ii) $\sqrt{n}\left(\hat\theta-\theta_0\right)$ is asymptotically normal with mean equal to zero and variance equal to $J^{-2}(\theta_0)\,K(\theta_0)$, where $J(\theta_0)$ and $K(\theta_0)$, under the assumption that the true distribution g belongs to the parametric family $\{f_\theta\}$ and θ_0 being the true value of the parameter, are given by:

$$J(\theta_0) = \int\left[u_{\theta_0}(z)\right]^2 f_{\theta_0}^{1+a}(z)\,dz$$

and

$$K\left(\theta_0\right) = \int \left[u_{\theta_0}\left(z\right)\right]^2 f_{\theta_0}^{1+2a}\left(z\right) dz - \xi^2 \tag{3.2.10}$$

where $\xi = \int u_{\theta_0}\left(z\right) f_{\theta_0}^{1+a}\left(z\right) dz$.

It is easy to see that by the weak law of large numbers, as $n \to \infty$, we have:

$$\left[\frac{\partial Q_\theta}{\partial \theta}\right]_{\theta=\theta_0} \xrightarrow{P} \left[\frac{\partial W_\theta}{\partial \theta}\right]_{\theta=\theta_0} \tag{3.2.11}$$

and

$$\left[\frac{\partial^2 Q_\theta}{\partial \theta^2}\right]_{\theta=\theta_0} \xrightarrow{P} \left[\frac{\partial^2 W_\theta}{\partial \theta^2}\right]_{\theta=\theta_0}. \tag{3.2.12}$$

The consistency of $\hat{\theta}$, the continuity of $J(\theta)$, expressions (3.2.8), (3.2.11) and (3.2.12) and a Taylor expansion of Q_θ around the point $\hat{\theta}$ can be used to evaluate the expectations of Q_θ and $W_{\hat{\theta}}$:

Theorem 3.2.3. *The expectation of* Q_θ *evaluated at the true point* θ_0 *is given by*

$$EQ_{\theta_0} = EQ_{\hat{\theta}} + \frac{a+1}{2} E\left[\left(\theta_0 - \hat{\theta}\right)^2 J(\theta_0)\right] + ER_n$$

and the expected overall discrepancy evaluated at $\hat{\theta}$ *is given by*

$$EW_{\hat{\theta}} = E\Big\{Q_{\hat{\theta}} + (a+1)\left(\hat{\theta} - \theta_0\right)^2 J(\theta_0) + R_n\Big\}$$

where R_n *and* $J(\theta_0)$ *as in Theorem 3.2.1.*

Proof. Since $\hat{\theta} \to \theta_0$ as $n \to \infty$, equations (3.2.4), (3.2.5), and (3.2.11), and under the assumption that the true distribution g belongs to the parametric family $\{f_\theta\}$ we have:

$$\left[\frac{\partial Q_\theta}{\partial \theta}\right]_{\theta=\hat{\theta}} \to 0$$

and

$$\left[\frac{\partial^2 Q_\theta}{\partial \theta^2}\right]_{\theta=\hat{\theta}} \to (a+1)\, J(\hat{\theta})$$

so that for large n we have for a Taylor expansion of Q_{θ_0} around the estimator $\hat{\theta}$, the following approximation:

$$Q_{\theta_0} = Q_{\hat{\theta}} + \frac{a+1}{2}\left(\theta_0 - \hat{\theta}\right)^2 J(\hat{\theta}) + o((\hat{\theta} - \theta_0)^2).$$

By the continuity of $J(\theta)$ we assert the first part of the theorem. For the second part observe that

$$E\left(Q_\theta \left|\theta = \theta_0\right.\right) = EQ_{\hat{\theta}} + \frac{a+1}{2}E\left[\left(\hat{\theta} - \theta_0\right)^2 J(\theta_0)\right] + ER_n \equiv W_{\theta_0}.$$

By combining the first part of the theorem and Theorem 3.2.1 we obtain the unbiasedness of the estimator of the expected overall discrepancy $EW_{\hat{\theta}}$. ■

3.2.3 The construction of the Divergence Information Criterion

Before the construction of the new criterion, the results of the previous two subsections will be extended to the multivariate case. This extension is possible since Theorem 3.2.2 holds for a $p-$dimensional parameter space Θ, $p \geq 1$ (Basu et *al.*, 1998). Indeed, in this case and under the same assumptions as those stated in Theorem 3.2.2 the $p-$dimensional estimator $\hat{\theta} = \left(\hat{\theta}_1, ..., \hat{\theta}_p\right)'$ is consistent for $\theta_0 = (\theta_{01}, ..., \theta_{0p})'$ and $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically multivariate normal with vector mean $\mathbf{0}$ and variance-covariance matrix $J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0)$ where

$$J\left(\theta_0\right) = \int u_{\theta_0}\left(z\right) u'_{\theta_0}\left(z\right) f_{\theta_0}^{1+a}\left(z\right) dz$$

and

$$K\left(\theta_0\right) = \int u_{\theta_0}\left(z\right) u'_{\theta_0}\left(z\right) f_{\theta_0}^{1+2a}\left(z\right) dz - \xi\xi', \qquad (3.2.13)$$

$\xi = \int u_{\theta_0}\left(z\right) f_{\theta_0}^{1+a}\left(z\right) dz$ and $u_\theta\left(z\right) = \frac{\partial}{\partial\theta}\left(\log\left(f_\theta\left(z\right)\right)\right)$.

As a result, for a $p-$dimensional parameter θ, we can see that (3.2.6) at $\theta = \hat{\theta}$ takes the form

$$EW_{\hat{\theta}} = W_{\theta_0} + \frac{(a+1)}{2}E\left[\left(\hat{\theta} - \theta_0\right)' J(\theta_0)\left(\hat{\theta} - \theta_0\right)\right] + E\Big\{o(||\hat{\theta} - \theta_0||^2)\Big\}. \qquad (3.2.14)$$

Similarly, the unbiasedness property of Theorem 3.2.3 takes the form:

$$EW_{\hat{\theta}} = E\Big\{Q_{\hat{\theta}} + (a+1)\left(\hat{\theta} - \theta_0\right)' J\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) + o(||\hat{\theta} - \theta_0||^2)\Big\}. \qquad (3.2.15)$$

Consider now the case that the candidate model f_θ comes from the family of the multivariate normal distribution where θ is the mean vector and $\hat{\theta}$ is obtained by minimizing 3.1.1 (Basu method). Then, it can be shown that (see Basu et *al.* (1998)),

$$J\left(\theta_0\right) = \left(2\pi\right)^{-\frac{a}{2}}\left(1 + a\right)^{-\left(1+\frac{p}{2}\right)}\Sigma^{-\left(1+\frac{a}{2}\right)}$$

and

$$Var\left(\hat{\theta}\right) = \left(1 + \frac{a^2}{1+2a}\right)^{1+\frac{p}{2}} \Sigma$$

so that

$$J\left(\theta_0\right) = (2\pi)^{-\frac{a}{2}} \left(\frac{1+a}{1+2a}\right)^{1+\frac{p}{2}} \Sigma^{-\frac{a}{2}} \left[Var\left(\hat{\theta}\right)\right]^{-1},$$

where Σ is the p x p asymptotic covariance matrix of the maximum likelihood estimator of the p - dimensional parameter θ_0.

Taking now into consideration the fact that

$$n \cdot o((\hat{\theta} - \theta_0)^2) = o_P(1)$$

since $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normal, we have that

$$n\left(\hat{\theta} - \theta_0\right)' \Sigma^{-\frac{a}{2}} \left[Var\left(\hat{\theta}\right)\right]^{-1} \left(\hat{\theta} - \theta_0\right) \tag{3.2.16}$$

has approximately a $\mathcal{X}_p^2$ distribution for a small. Then, the Divergence Information Criterion defined as the asymptotically unbiased estimator of $EW_{\hat{\theta}}$ is introduced in the theorem below.

Theorem 3.2.4. *Assume that the candidate model comes from the family of the multivariate normal distribution with θ the mean vector and $\hat{\theta}$ the estimator obtained by minimizing 3.1.1. An asymptotically unbiased estimator of $n-$times the expected overall discrepancy evaluated at $\hat{\theta}$ is given by*

$$DIC = nQ_{\hat{\theta}} + (a+1)\,(2\pi)^{-\frac{a}{2}} \left(\frac{1+a}{1+2a}\right)^{1+\frac{p}{2}} p. \tag{3.2.17}$$

The DIC criterion as it has been derived in the above theorem uses as an estimator of the unknown parameter the estimator obtained by minimizing (3.1.1) (Basu method). As it was mentioned earlier, the researcher may alternatively choose to use the maximum likelihood method (MLE method) in which case the correction term is adjusted accordingly. Indeed, in this case

$$\begin{aligned} J\left(\theta_0\right) &= (2\pi)^{-\frac{a}{2}} (1+a)^{-\left(1+\frac{p}{2}\right)} \Sigma^{-\left(1+\frac{a}{2}\right)} \\ &= (2\pi)^{-\frac{a}{2}} (1+a)^{-\left(1+\frac{p}{2}\right)} \Sigma^{-\frac{a}{2}} \left[Var\left(\hat{\theta}\right)\right]^{-1} \end{aligned}$$

since $Var\left(\hat{\theta}\right) = \Sigma$ is the covariance matrix of the maximum likelihood estimator. Using (3.2.15) and the fact that (3.2.16) follows again approximately a $\mathcal{X}_p^2$ distribution it is easy to see that the adjusted DIC is given by

$$DIC_{MLE} = nQ_{\hat{\theta}} + (2\pi)^{-\frac{a}{2}} (1+a)^{-\frac{p}{2}} p. \tag{3.2.18}$$

By comparing the correction terms of DIC and DIC_{MLE} we observe that they are similar in the sense that for small a

$$(1+a)\left(\frac{1+a}{1+2a}\right)^{1+\frac{p}{2}} \simeq (1+a)^{-\frac{p}{2}} < 1.$$

In order to put into the proposed criterion some extra penalty for too large models (models with large number of parameters) we can replace the above term(s) by a (common) quantity larger than 1. Observe that for small values of a the denominator of the left hand side of the above expression can be assumed to be close to 1 and therefore it can be disregarded. As a result both of the above terms can be replaced in DIC and DIC_{MLE} by the remaining part of the expression on the left hand side, namely

$$(1+a)^{2+\frac{p}{2}}.$$

Observe that the above quantity is now larger than 1 so that the penalty term of the criterion will be larger for large values of p. Both criteria are adjusted accordingly and in fact now, they are both given by the same corrected formula (although $\hat{\theta}$ is obtained by different estimating methods), namely

$$DIC_c = nQ_{\hat{\theta}} + (2\pi)^{-\frac{a}{2}} (1+a)^{2+\frac{p}{2}} p. \tag{3.2.19}$$

The MLE method and the associated DIC_{MLE} and DIC_c have a number of advantages. In particular, the MLE method is computationally faster than the Basu method. This is due to the fact that the MLE method is given in closed form as opposed to the Basu method which is not in closed form and as a result we rely on a numerical method to obtain the desired estimator. Such numerical methods are usually associated with errors which may not be controllable, a feature that makes such methods unattractive. As a consequence, the MLE method is more accurate than the Basu method and at the same time satisfies the standard properties required by such estimators, namely the consistency and the asymptotic normality. The practical implications of these two forms of the DIC criterion become evident in Chapter 5 were simulations are performed.

Observe that the DIC criterion consists of two terms. The first term, $Q_{\hat{\theta}}$, is a biased estimator of the expected overall discrepancy. As a result, if we choose the model with the smallest estimator of the expected overall discrepancy we may end up with a selection with an unnecessarily large number of covariates. The estimator becomes asymptotically unbiased by introducing the appropriate correction term according to the estimating method used. The correction term could be viewed also as a penalty term for too large dimension p.

3.3 Lower Bound of the MSE of Prediction

One of the main issues in model selection is the notion of asymptotic efficiency [Shibata, 1980; 1981]. The asymptotic efficiency deals with the selection of a model with finitely many variables that provides the best possible approximation of the true model with infinity many variables with respect to the mean squared error (MSE) of prediction. The issue of asymptotic efficiency is of great interest whenever the object of the analysis is a model selection that yields a good inference. Here we provide a lower bound for the mean squared error of prediction. In particular we show that the MSE of prediction of DIC is never below the so called Average Mean Squared Error (Average MSE) of prediction. For the evaluation of the MSE the original set of n observations are used for the estimation of the parameters and the one-step ahead prediction is used for measuring the accuracy of the selection. Following Shibata's assumption [Shibata, 1981] infinitely many independent variables are assumed so that the design matrix $\mathbf{X}$ is a $n \times \infty$ matrix.

Let $\mathbf{X}$ be the design matrix of the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots)'$ the vector of unknown coefficients, $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I)$ the error sequence and I the infinite dimensional identity matrix.

Let

$$V\left(j\right)=\left\{c\left(j\right),\ such\ that\ c\left(j\right)=\left(c_0,0,...,c_{j_1},0,...,c_{j_{k_j}},0,...\right)'\right\}$$

be the subspace that contains the k_j+1 parameters involved in the model and let

$$\boldsymbol{\beta}^{(\mathbf{n})}=\left(\beta_0,0,...,\beta_{j_1},0,...,\beta_{j_{k_j}},0,...\right)'$$

be the projection of $\boldsymbol{\beta}$ on $V\left(j\right)$.

The prediction $\hat{\mathbf{Y}}=(\hat{Y}_1,\ldots,\hat{Y}_n)'$ is given by

$$\hat{\mathbf{Y}}=\mathbf{X_j}\hat{\boldsymbol{\beta}},$$

where the estimator of $\boldsymbol{\beta}^{(\mathbf{n})}$ obtained through a set of observations $(X_{ij_1},\ldots,X_{ij_{k_j}},Y_i)$, $i=1,2,\ldots,n$ is denoted by

$$\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,0,...,\hat{\beta}_{j_1},0,...,\hat{\beta}_{j_2},0,...,\hat{\beta}_{j_{k_j}},0,...\right)'.$$

Observe that the design matrix $\mathbf{X_j}$ is a $n\times\infty$ matrix where only the columns $j_1,\ldots,j_{k_j}$ have entries different than zero.

The mean squared error (MSE) of prediction (up to a constant) and the average MSE of prediction are defined respectively by

$$S_n(j)=\mathrm{E}\left[\left(\hat{\mathbf{Y}}-\mathbf{Y}\left|\mathbf{X_j}\right.\right)'\left(\hat{\mathbf{Y}}-\mathbf{Y}\left|\mathbf{X_j}\right.\right)\right]-n\sigma^2$$

and

$$L_n(j)\equiv\mathrm{E}\left(S_n(j)\right).$$

We will prove now that the above two quantities take the form given in the following Lemma.

Lemma 3.3.1. *Under the notation and conditions of this section we have that*

$$S_n(j)=\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2_{\mathbf{M_n(j)}}$$

and

$$L_n(j)=\mathrm{E}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2_{\mathbf{M_n(j)}},$$

where $M_n\left(j\right)=X_j'X_j$ *and* $\left\|A\right\|_R^2=A'RA$.

Proof. It is easy to see that

$$\begin{aligned}
E(\hat{\mathbf{Y}} - \mathbf{Y}\,|\mathbf{X_j})'(\hat{\mathbf{Y}} - \mathbf{Y}\,|\mathbf{X_j}) &= E(\mathbf{X_j}\hat{\boldsymbol{\beta}} - \mathbf{X_j}\boldsymbol{\beta} - \boldsymbol{\varepsilon}|\mathbf{X_J})'(\mathbf{X_j}\hat{\boldsymbol{\beta}} - \mathbf{X_j}\boldsymbol{\beta} - \boldsymbol{\varepsilon}|\mathbf{X_J}) \\
&= E\left(\mathbf{X_j}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}|\mathbf{X_J}\right)'\left(\mathbf{X_j}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{\varepsilon}|\mathbf{X_J}\right) \\
&= E\Big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X_j}'\mathbf{X_j}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\quad + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}'\mathbf{X_j}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)|\mathbf{X_J}\Big) \\
&= \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\mathbf{X_j}'\mathbf{X_j}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + n\sigma^2 \\
&= \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2_{\mathbf{M_n(j)}} + n\sigma^2.
\end{aligned}$$

The results follow immediately. ■

The Lemma below provides a lower bound for the MSE of prediction. In particular, we show that $S_n(j)$ is asymptotically never below the quantity

$$L_n\left(j^*\right) = \min_j L_n(j).$$

Lemma 3.3.2. *Let $L_n\left(j^*\right) = \min_j L_n(j)$. Assume also that for $0 < \delta < 1$*

$$\lim_{n\to\infty}\sum_j\left[\left(1 - \delta\omega_n\left(j\right)\right)\exp\left(\delta\omega_n\left(j\right)\right)\right]^{\frac{k_j+1}{2}} = 0,$$

where

$$\omega_n\left(j\right) = \frac{L_n\left(j\right)}{(k_j+1)g\left(a, k_j+1\right)\sigma^2}$$

and $g(a,m) = (1 + a^2/(1+2a))^{\frac{m}{2}+1}$. Then, for every $0 < \delta < 1$

$$\lim_{n\to\infty} P\left[\frac{S_n\left(j\right)}{L_n\left(j^*\right)} > 1 - \delta\right] = 1.$$

Proof. For every $0 < \delta < 1$ and for every j and using the fact that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2_{\mathbf{M_n(j)}} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(n)}\|^2_{\mathbf{M_n(j)}} + \|\boldsymbol{\beta}^{(n)} - \boldsymbol{\beta}\|^2_{\mathbf{M_n(j)}}$$

we have

$$\begin{aligned}
P\left[\frac{S_n(j)}{L_n(j^*)} \leq 1-\delta\right] &\leq P\left[\frac{S_n(j)}{L_n(j)} \leq 1-\delta\right] \\
&\leq \sum_j P\left[\frac{\left\|\hat{\beta}-\beta\right\|^2_{\mathbf{M_n(j)}}}{L_n(j)} \leq 1-\delta\right] \\
&= \sum_j P\left[\frac{\left\|\hat{\beta}-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} + \left\|\beta^{(\mathbf{n})}-\beta\right\|^2_{\mathbf{M_n(j)}}}{L_n(j)} \leq 1-\delta\right] \\
&= \sum_j P\left[\frac{\left\|\hat{\beta}-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} + \left\|\beta^{(\mathbf{n})}-\beta\right\|^2_{\mathbf{M_n(j)}}}{L_n(j)} \leq 1-\delta\right] \\
&= \sum_j P\left[\left\|\hat{\beta}-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} \leq (1-\delta)\,L_n(j) - \left\|\beta^{(\mathbf{n})}-\beta\right\|^2_{\mathbf{M_n(j)}}\right]
\end{aligned} \tag{3.3.1}$$

By Theorem 3.2.2 the limiting covariance matrix of $n^{1/2}\hat{\theta}$ is a multivariate normal random variable

$$N_p\left(\theta_0, g(a,p)\Sigma\right),$$

where

$$g(\alpha,p) = \left(1+\frac{\alpha^2}{1+2\alpha}\right)^{p/2+1}.$$

Then, in this case we have

$$\begin{aligned}
\left\|\hat{\beta}-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} &= \left(\hat{\beta}-\beta^{(n)}\right)'\left\{\sigma^2 g\left(\alpha,k_j+1\right)\mathbf{M_n}\left(\mathbf{j}\right)\right\}^{-1}\left(\hat{\beta}-\beta^{(n)}\right)\sigma^2 g\left(\alpha,k_j+1\right) \\
&\sim \sigma^2 g\left(\alpha,k_j+1\right)\mathcal{X}^2_{k_j+1}
\end{aligned} \tag{3.3.2}$$

and

$$\begin{aligned}
L_n(j) &= E\left\|\hat{\beta}-\beta\right\|^2_{\mathbf{M_n(j)}} \\
&= \left\|\beta-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} + E\left\|\hat{\beta}-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} \\
&= \left\|\beta-\beta^{(n)}\right\|^2_{\mathbf{M_n(j)}} + \left(k_j+1\right)g\left(\alpha,k_j+1\right)\sigma^2.
\end{aligned}$$

Using (3.3.2) we have that (3.3.1) is bounded by

$$\begin{aligned}
&\sum_j P\left[\mathcal{X}^2_{k_j+1} \leq (k_j+1) - \delta(k_j+1)\omega_n(j)\right] \\
&\leq \sum_j \left[\exp\left(\delta\omega_n(j)\right)\left(1-\delta\omega_n(j)\right)\right]^{\frac{k_j+1}{2}}
\end{aligned}$$

where the last inequality follows from the fact that for $k > \delta$ [see Shibata, 1981]

$$P\left[\mathcal{X}_k^2 \leq k - \delta\right] \leq \exp\left(\frac{\delta}{2}\right)\left(1 - k^{-1}\delta\right)^{\frac{k}{2}} \leq \exp\left(\frac{-\delta^2}{4k}\right). \tag{3.3.3}$$

By the assumption of the theorem we get

$$\begin{aligned} \lim_{n\to\infty} P\left[\frac{S_n(j)}{L_n(j^*)} \leq 1 - \delta\right] &= 0 \Rightarrow \\ \lim_{n\to\infty} P\left[\frac{S_n(j)}{L_n(j^*)} > 1 - \delta\right] &= 1. \end{aligned}$$

■

Chapter 4

Goodness of Fit Statistics

4.1 Introduction

It is important to state that model selection criteria are considered as statistics which could be used for inferential purposes. More specifically, any model selection criterion can be used for making a selection among competing hypotheses. Indeed, consider a set of candidate models each of which may be the underlying process which the available data came from. In that sense, each candidate model forms a hypothesis. Then, each of the competing hypotheses is fitted to the data and the value of the model selection criterion is computed. In such cases we select among the competing hypotheses, the one for which the model selection criterion is minimized (for details see Sakamoto et. al, 1986, Chapter 3).

One of the drawbacks of such a procedure is associated with the fact that the statistical significance of any difference observed in the values of the criterion for the competing hypotheses cannot be verified or evaluated. As a result, there is a need to provide a formal hypothesis testing procedure using the measure on which the model selection criterion is based.

The statistical analysis and in particular the testing of models for discrete multivariate data has been given considerable attention during the last 30 years. The books of Cox (1970), Gokhale and Kullback (1978), Agresti (1984) and Cressie and Read (1988) are focusing on various aspects of model development. The usual practice is that the adequacy of a model can be tested by one of the traditional goodness-of-fit tests, namely the Pearson's X^2 or the loglikelihood ratio test. Note that both of these tests are special cases of the Cressie and Read measure of divergence introduced in (1.5.5). Indeed in a discrete setting and for $\lambda = 1$ the Cressie and Read measure reduces to

$$\sum_{i=1}^{m} \frac{(p_i - q_i)^2}{q_i}$$

which multiplied by $2n$ is the Pearson's X^2 test where p_i plays the role of the observed frequency and q_i the role of the expected one. Furthermore, the loglikelihood ratio test statistic (also known as Kullback-Leibler measure, see Lemma 2.2.3)

$$2n \sum_{i=1}^{m} p_i \log\left(\frac{p_i}{q_i}\right)$$

can be deduced from the Cressie and Read measure for $\lambda \to 0$.

In this Chapter we focus on a discrete setting and provide initially the distributional properties of the estimator of the general BHHJ family of measures which is shown to be weakly consistent. These results are then used for establishing in Section 4.3 a goodness of fit test for multinomial distributions based on the general BHHJ family of divergence measures.

4.2 Distributional Properties

Definition 4.2.1. *Let f be a continuous, convex, homogeneous function defined on the set*

$$S_k = \{(s_1, s_2) : 0 < s_i < \infty,\ i = 1, 2\},$$

with continuous derivatives of second order. Then the f-dissimilarity is defined to be

$$d_a = d_f(Q, P) = \sum_{j=1}^{m} f(p_j, q_j)$$

where $p_j, q_j, j = 1, ..., m$ are the parameters from the multinomial distributions $M(N_p, P)$, $P = (p_1, p_2, ..., p_m)$ and $M(N_q, Q)$, $Q = (q_1, q_2, ..., q_m)$.

For different functions f we have specific dissimilarity measures. For example for

$$f(p, q) = q^{1+a}\Phi(p/q)$$

we have the general BHHJ family of measures for a general function Φ while for Φ as in (1.5.4) we have the discrete BHHJ measure and for $\Phi = \phi$ with $\alpha = 0$ we have the Csiszár's measure. Observe that the estimator of d_a is

$$\hat{d}_a = d_f\left(\hat{Q}, \hat{P}\right) = \sum_{j=1}^{m} f\left(\hat{p}_j, \hat{q}_j\right).$$

For the general BHHJ family of measures the estimator of the f-dissimilarity is given by

$$\hat{d}_a = \sum_{j=1}^{m} \hat{q}_j^{1+a} \Phi\left(\frac{\hat{p}_j}{\hat{q}_j}\right) \tag{4.2.1}$$

where $\hat{p}_j = \frac{x_j}{N_p}$, $\hat{q}_j = \frac{y_j}{N_q}$, j = 1,...,m, and $\mathbf{X} = (x_1, ..., x_m)$, $\mathbf{Y} = (y_1, ..., y_m)$ are random observations from $M(N_p, P)$ and $M(N_q, Q)$.

Observe that in case one of the two independent distributions is known then the obvious notation applies, namely

$$\hat{d}_a = d_f\left(Q, \hat{P}\right) = \sum_{j=1}^{m} f\left(\hat{p}_j, q_j\right)$$

if Q is known and

$$\hat{d}_a = d_f\left(\hat{Q}, P\right) = \sum_{j=1}^{m} f\left(p_j, \hat{q}_j\right)$$

if P is known.

Distributional properties and goodness of fit tests using measures of divergence such as Csiszár's have been extensively investigated [Zografos et *al.*, 1990; Morales et *al.*, 1997; Pardo, 1999 etc.]. In what follows we establish the distributional properties of (4.2.1) and provide approximations of the moments of the estimator of the general BHHJ family of measures.

Theorem 4.2.1. *Given two independent random observations $X = (x_1, x_2, ..., x_m)$ and $Y = (y_1, y_2, ..., y_m)$ from multinomial distributions $M(N_p, P)$, $P = (p_1, p_2, ..., p_m)$ and $M(N_q, Q)$, $Q = (q_1, q_2, ..., q_m)$ the expected value of $\hat{d}_a$ is approximately equal to:*

$$E\left(\hat{d}_a\right) = d_a + \frac{1}{2N_p}\sum_{j=1}^{m} p_j\left(1-p_j\right) q_j^{a-1}\Phi''\left(\frac{p_j}{q_j}\right) + \frac{1}{2N_q}\sum_{j=1}^{m}\left(1-q_j\right) q_j^{a}\psi(p,q,a) +$$

$$+o(N_p^{-1}) + o(N_q^{-1})$$

where $d_a = \sum\limits_{j=1}^{m} q_j^{1+a}\Phi\left(\frac{p_j}{q_j}\right)$,$\Phi$ *any function such that* $\Phi'(1) = 0$ *and* $\Phi''(1) \neq 0$ *and*

$$\psi(p,q,a) = a(a+1)\Phi\left(\frac{p_j}{q_j}\right) - 2a\frac{p_j}{q_j}\Phi'\left(\frac{p_j}{q_j}\right) + \frac{p_j^2}{q_j^2}\Phi''\left(\frac{p_j}{q_j}\right).$$

Proof. By Zografos [1987, Theorem 4.2.1, p. 148] we deduce that the expected value of the estimator of the f-dissimilarity is approximately equal to:

$$\begin{aligned} E\left(\hat{d}_a\right) = d_f(Q,P) + \frac{1}{2N_p}\sum_{j=1}^{m} p_j\left(1-p_j\right)\left[f''_{(p_j)}\left(p_j,q_j\right)\right] + o(\|\hat{P}-P\|^2) \\ +\frac{1}{2N_q}\sum_{j=1}^{m} q_j\left(1-q_j\right)\left[f''_{(q_j)}\left(p_j,q_j\right)\right] + o(\|\hat{Q}-Q\|^2) \quad (4.2.2) \end{aligned}$$

where $\hat{Q} = (\hat{q}_1,\ldots,\hat{q}_m)'$ and $\hat{P} = (\hat{p}_1,\ldots,\hat{p}_m)'$ the estimators of Q and P, $f : \Re^2 \to \Re$ a function with continuous second order partial derivatives in every point of an open subset of $\Re^2$ and

$$f''_{(s_i)}\left(s_1,s_2\right) = \frac{\partial^2}{\partial s_i^2} f\left(s_1,s_2\right),\ i = 1,2.$$

Take

$$f\left(p,q\right) = q^{1+a}\Phi\left(\frac{p}{q}\right),$$

with $\Phi\left(\cdot\right)$ as in the statement of the theorem. Then using

$$f''_{(p)}\left(p,q\right) = q^{a-1}\Phi''\left(\frac{p}{q}\right) \text{ and } f''_{(q)}\left(p,q\right) = q^{a-1}\psi(p,q,a)$$

in (4.2.2) and the facts that $N_q \cdot o(\|\hat{Q}-Q\|^2) = N_p \cdot o(\|\hat{P}-P\|^2) = o_P(1)$ we obtain the desired result. ■

Theorem 4.2.2. *Let two independent random observations* $X_i = (x_{i1},...,x_{im})$, *from multinomial distributions* $M\left(N_{X_i},P_i\right)$, *where* $P_i = \left(p_{i1},...,p_{im}\right)$, $i = 1,2$ *and another two independent random observations* $Y_i = (y_{i1},...,y_{im})$, *from multinomial distributions* $M\left(N_{Y_i},Q_i\right)$ *where* $Q_i = \left(q_{i1},...,q_{im}\right)$, $i = 1,2$. *Then the covariance of the estimators of the* f*-dissimilarities*

$$\hat{d}_{1a_1} = \sum_{j=1}^{m}\hat{p}_{2j}^{1+a_1}\Phi_1\left(\frac{\hat{p}_{1j}}{\hat{p}_{2j}}\right) \text{ and } \hat{d}_{2a_2} = \sum_{j=1}^{m}\hat{q}_{2j}^{1+a_2}\Phi_2\left(\frac{\hat{q}_{1j}}{\hat{q}_{2j}}\right),$$

with $\Phi_i(u)$, $i = 1,2$ *any functions such that* $\Phi_i'(1) = 0$ *and* $\Phi_i''(1) \neq 0$, $i = 1,2$, *is* *asymptotically equal to:*

$$\begin{aligned} Cov\left(\hat{d}_{1a_1}, \hat{d}_{2a_2}\right) = \sum_{i,j} p_{2i}^{a_1} q_{2j}^{a_2} \Big\{ & \Phi_1'\left(\frac{p_{1i}}{p_{2i}}\right)\Phi_2'\left(\frac{q_{1j}}{q_{2j}}\right) Cov\left(\hat{p}_{1i}, \hat{q}_{1j}\right) \\ & + \Phi_1'\left(\frac{p_{1i}}{p_{2i}}\right)\left[(1+a_2)\Phi_2\left(\frac{q_{1j}}{q_{2j}}\right) - \frac{q_{1j}}{q_{2j}}\Phi_2'\left(\frac{q_{1j}}{q_{2j}}\right)\right] Cov\left(\hat{p}_{1i}, \hat{q}_{2j}\right) \\ & + \Phi_2'\left(\frac{q_{1j}}{q_{2j}}\right)\left[(1+a_1)\Phi_1\left(\frac{p_{1i}}{p_{2i}}\right) - \frac{p_{1i}}{p_{2i}}\Phi_1'\left(\frac{p_{1i}}{p_{2i}}\right)\right] Cov\left(\hat{p}_{2i}, \hat{q}_{1j}\right) \\ & + \left[(1+a_1)\Phi_1\left(\frac{p_{1i}}{p_{2i}}\right) - \frac{p_{1i}}{p_{2i}}\Phi_1'\left(\frac{p_{1i}}{p_{2i}}\right)\right] \times \\ & \times \left[(1+a_2)\Phi_2\left(\frac{q_{1j}}{q_{2j}}\right) - \frac{q_{1j}}{q_{2j}}\Phi_2'\left(\frac{q_{1j}}{q_{2j}}\right)\right] Cov\left(\hat{p}_{2i}, \hat{q}_{2j}\right)\Big\} + R_N, \end{aligned}$$

where $R_N = o(N^{-1})$ *assuming that* $N_{X_i} = N_{Y_i} = N$, $\forall i$.

Proof. The first order Taylor expansions of the estimators

$$\hat{d}_{f_1} = \sum_{j=1}^{m} f_1\left(\hat{p}_{1j}, \hat{p}_{2j}\right) \text{ and } \hat{d}_{f_2} = \sum_{j=1}^{m} f_2\left(\hat{q}_{1j}, \hat{q}_{2j}\right)$$

of the dissimilarities

$$d_{f_1}(Q,P) = \sum_{j=1}^{m} f_1\left(p_{1j}, p_{2j}\right) \text{ and } d_{f_2}(Q,P) = \sum_{j=1}^{m} f_2\left(q_{1j}, q_{2j}\right)$$

are given by

$$\hat{d}_{f_1} = d_{f_1}(Q,P) + \sum_{i=1}^{2}\sum_{j=1}^{m}(\hat{p}_{ij} - p_{ij}) f_{1,(i)}\left(p_{1j}, p_{2j}\right) + o(||\hat{P} - P||)$$

and

$$\hat{d}_{f_2} = d_{f_2}(Q,P) + \sum_{i=1}^{2}\sum_{j=1}^{m}(\hat{q}_{ij} - q_{ij}) f_{2,(i)}\left(q_{1j}, q_{2j}\right) + o(||\hat{Q} - Q||),$$

where $\hat{p}_{ij} = \frac{x_{ij}}{N_{X_i}}$, $\hat{q}_{ij} = \frac{y_{ij}}{N_{Y_i}}$, $i = 1, 2$, $j = 1, 2, ..., m$, and

$$f_{k,(s)}\left(w_1, w_2\right) = \frac{\partial}{\partial w_s} f_k\left(w_1, w_2\right), k, s = 1, 2.$$

By Zografos [1987, Theorem 4.2.2, p. 150] we have that the covariance of the above estimators is equal to:

$$Cov\left(\hat{d}_{f_1}, \hat{d}_{f_2}\right) = \sum_{i,v=1}^{2}\sum_{j,\mu=1}^{m} f_{1,(i)}\left(p_{1j}, p_{2j}\right) f_{2,(v)}\left(q_{1\mu}, q_{2\mu}\right) Cov\left(\hat{p}_{ij}, \hat{q}_{v\mu}\right) + R_N,$$

where R_N the remainder term. The theorem is derived by using

$$f_1\left(p_1, p_2\right) = p_2^{1+a_1}\Phi_1\left(\frac{p_1}{p_2}\right), \quad f_2\left(q_1, q_2\right) = q_2^{1+a_2}\Phi_2\left(\frac{q_1}{q_2}\right),$$

$$f_{k,(1)}\left(w_1,w_2\right)=\frac{\partial}{\partial w_1}f_k\left(w_1,w_2\right)=w_{_2}^{a}\Phi_k'\left(\frac{w_1}{w_2}\right),\ k=1,2,$$

$$\begin{aligned}f_{k,(2)}\left(w_1,w_2\right)&=\frac{\partial}{\partial w_2}f_k\left(w_1,w_2\right)\\&=w_{_2}^{a}\left[\left(1+a_k\right)\Phi_k\left(\frac{w_1}{w_2}\right)-\frac{w_1}{w_2}\Phi_k'\left(\frac{w_1}{w_2}\right)\right],k=\ 1,2.\end{aligned}$$

with $\Phi_1\left(u\right)$ and $\Phi_2\left(u\right)$ as in the statement of the theorem.

Without loss of generality assume that $N_{X_i}=N_{Y_i}=N$, $\forall i$. It is easily seen that with the use of the relations

$$o(\sqrt{N})o(\sqrt{N})=o(N^{-1}),\ o(N^{-1})+o(N^{-1})=o(N^{-1})$$

$$||\hat{P}-P||=||\hat{Q}-Q||=O_P(N^{-1/2})\text{ and }o(O_P(N^{-1/2}))=o(N^{-1/2}),$$

the remainder term turns out to be equal to $o(N^{-1})$. ■

Corollary 4.2.2. *Given two independent random observations $X=(x_1,...,x_m)$ and $Y=(y_1,...,y_m)$ from multinomial distributions $M\left(N_p,P\right),P=(p_1,...,p_m)$ and $M\left(N_q,Q\right),Q=(q_1,...,q_m)$, the variance of the estimator $\hat{d}_a$ of the f-dissimilarity d_a is asymptotically equal to:*

$$\begin{aligned}Var\left(\hat{d}_a\right)=&\frac{1}{N_p}\sum_{j=1}^{m}p_jq_j^{2a}\left[\Phi'\left(\frac{p_j}{q_j}\right)\right]^2-\frac{1}{N_p}\left[\sum_{j=1}^{m}p_jq_j^{a}\Phi'\left(\frac{p_j}{q_j}\right)\right]^2\\&+\frac{1}{N_q}\sum_{j=1}^{m}q_jq_j^{2a}\left[(1+a)\Phi\left(\frac{p_j}{q_j}\right)-\frac{p_j}{q_j}\Phi'\left(\frac{p_j}{q_j}\right)\right]^2\\&-\frac{1}{N_q}\left[\sum_{j=1}^{m}q_jq_j^{a}\Big((1+a)\Phi\left(\frac{p_j}{q_j}\right)-\frac{p_j}{q_j}\Phi'\left(\frac{p_j}{q_j}\right)\Big)\right]^2+o(N_p^{-1})+o(N_q^{-1}).\end{aligned}$$

Proof. It follows immediately from the previous theorem since for

$$p_{1j}=q_{1j}\equiv p_j,\ \ p_{2j}=q_{2j}\equiv q_j,\ \Phi_1=\Phi_2\text{ and }a_1=a_2$$

the covariance reduces to the variance of the estimator $\hat{d}_a$. ■

This section ends with the consistency property of the proposed estimator $\hat{d}_a$.

Corollary 4.2.3. *Let two independent random observations $X=(x_1,...,x_m)$ and $Y=(y_1,...,y_m)$ from multinomial distributions $M\left(N_p,P\right),P=(p_1,...,p_m)$ and $M\left(N_q,Q\right),Q=(q_1,...,q_m)$. Then the estimator $\hat{d}_a$ is a weakly consistent estimator of the f-dissimilarity d_a.*

PROOF. The result follows immediately from Theorem 4.2.1 and Corollary 4.2.2 since, as $N_q \to \infty$ and $N_p \to \infty$,

$$E\left(\hat{d}_a\right) \to d_a \; and \; Var\left(\hat{d}_a\right) \to 0.$$

■

4.3 Goodness of Fit Tests

If we have to examine whether the data $(n_1, n_2, ..., n_m)$ come from a multinomial distribution $M(N, P_0)$, *where* $P_0 = (p_{10}, p_{20}, ..., p_{m0})$ and $N = \sum_{i=1}^{m} n_i$, a well known test statistic is the chi-square goodness of fit test statistic. We define now for any function Φ such that $\Phi'(1) = 0$ and $\Phi''(1) \neq 0$, a new statistic for the above goodness of fit test:

$$\mathrm{X}_a^2 \equiv \frac{2N\left(\hat{d}_a - \Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a}\right)}{\Phi''(1)} \tag{4.3.1}$$

which for $\Phi(u)$ as in (1.5.4) constitutes the test statistic associated with the BHHJ divergence. Observe that for the purpose of goodness of fit tests we use

$$\hat{d}_a = \sum_{i=1}^{m} {q_i}^{1+a} \Phi\left(\frac{\hat{p}_i}{q_i}\right)$$

with $q_i = p_{i0}$.

In what follows we establish the asymptotic distribution of the estimator $\hat{d}_a$ (Corollary 4.3.1) and the test statistic (4.3.1) (Theorem 4.3.2).

Theorem 4.3.1. *Let* $g : \Re^k \to \Re$ *a function of the form*

$$g(x_1, x_2, ..., x_m) = \sum_{i=1}^{m} q_i^{1+a} \Phi(x_i/q_i),$$

with $\Phi(u)$ *any function such that* $\Phi'(1) = 0$ *and* $\Phi''(1) \neq 0$ *and* q_i *known. Then*

$$\sqrt{N}\left[g(\hat{p}_1, ..., \hat{p}_m) - g(p_1, ..., p_m)\right] \xrightarrow{L} N\left(0, \sigma_a^2\right)$$

where

$$\sigma_a^2 = \left\{\sum_{j=1}^{m} p_j \left[q_j^a \Phi'\left(\frac{p_j}{q_j}\right)\right]^2 - \left[\sum_{j=1}^{m} p_j q_j^a \Phi'\left(\frac{p_j}{q_j}\right)\right]^2\right\}$$

and $\hat{p}_i = \frac{x_i}{N}$, $i = 1, .., m$.

Proof. Since $X = (x_1, x_2, ..., x_m)$ is a random observation from the multinomial distribution $M(N, P)$, $P = (p_1, p_2, ..., p_m)$ and $\hat{p}_i = \frac{x_i}{N}, i = 1, .., m$ it follows that (see, e.g. Serfling, 1980, p. 108-109),

$$\sqrt{N}\left(\hat{p}_1 - p_1, \hat{p}_2 - p_2, ..., \hat{p}_m - p_m\right) \xrightarrow{L} N(0, \Sigma),$$

where the variance-covariance matrix is given by $\Sigma = [\sigma_{ij}]_{mxm}$,

$$\sigma_{ij} = \begin{cases} p_i(1-p_i), \ i = j \\ -p_i p_j, \ i \neq j \end{cases}$$

The theorem is derived by applying the well known Delta method to the case under investigation (for a similar result see Rao, 1973, p. 387) with

$$\sigma_a^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}\frac{\partial g}{\partial p_i}\frac{\partial g}{\partial p_j},$$

where

$$\frac{\partial g}{\partial p_k} = q_k^a \Phi'(p_k/q_k), \ k = 1, 2, \ldots, m.$$

Indeed, in this case we have

$$\begin{aligned}\sigma_a^2 &= \sum_{i=1}^{m} p_i(1-p_i)\left[q_i^a\Phi'\left(\frac{p_i}{q_i}\right)\right]^2 - \sum\sum\nolimits_{i\neq j} p_i p_j \left[q_i^a\Phi'\left(\frac{p_i}{q_i}\right)\right]\left[q_j^a\Phi'\left(\frac{p_j}{q_j}\right)\right] \\ &= \sum_{i=1}^{m} p_i\left[q_i^a\Phi'\left(\frac{p_i}{q_i}\right)\right]^2 - \sum_{i=1}^{m} p_i^2\left[q_i^a\Phi'\left(\frac{p_i}{q_i}\right)\right]^2 \\ &\qquad - \sum\sum\nolimits_{i\neq j} p_i p_j \left[q_i^a\Phi'\left(\frac{p_i}{q_i}\right)\right]\left[q_j^a\Phi'\left(\frac{p_j}{q_j}\right)\right]\end{aligned}$$

and the result is immediate. ■

Corollary 4.3.1. *Let d_a as in (1.5.6) and any function Φ such that $\Phi'(1) = 0$ and $\Phi''(1) \neq 0$ with $q_i \equiv p_{i0}$, $i = 1, \ldots, m$. Then*

$$\sqrt{N}\left[\hat{d}_a - d_a\right] \xrightarrow{L} N\left(0, \sigma_a^2\right)$$

where

$$\sigma_a^2 = \left\{\sum_{j=1}^{m} p_j \left[p_{j0}^a \Phi'\left(\frac{p_j}{p_{j0}}\right)\right]^2 - \left[\sum_{j=1}^{m} p_j p_{j0}^a \Phi'\left(\frac{p_j}{p_{j0}}\right)\right]^2\right\}$$

and

$$\hat{d}_a = \sum_{i=1}^{m} p_{i0}^{1+a}\Phi\left(\frac{\hat{p}_i}{p_{i0}}\right).$$

Proof. It follows immediately from the previous theorem. ■

We provide below the definition of the usual stochastic ordering which is used in Theorem 4.3.2 where the asymptotic distribution of the test statistic (4.3.1) under the null hypothesis $H_0 : p_i = p_{i0},\ i = 1, ..., m$ is established.

Definition 4.3.2. *Let X and Y continuous random variables with cdfs F and G. Let F^{-1} and G^{-1} the inverses and $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$ the corresponding survival functions. X is said to be smaller than Y in the usual stochastic order $X \prec_{st} Y$ if*

$$\bar{F}(x) \leq \bar{G}(x)\ \forall x \in R.$$

Also $X \prec_{st} Y$ iff $F^{-1}(p) \leq G^{-1}(p),\ p \in (0, 1)$.

Theorem 4.3.2. *Let $(n_1, ..., n_m) \sim M(N, P)$ with $P = (p_1, ..., p_m)$ and $p_i,\ i = 1, ..., m$ unknown parameters. Under the null hypothesis $H_0 : p_i = p_{i0},\ i = 1, ..., m$ we have:*

- $\left(\min\limits_i p_{i0}^a\right) \sum\limits_{i=1}^m \frac{N}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \prec_{\text{st}} \sum\limits_{i=1}^m \frac{Np_{i0}^a}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \prec_{\text{st}} \left(\max\limits_i p_{i0}^a\right) \sum\limits_{i=1}^m \frac{N}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2$
- $\mathrm{X}_a^2 - \sum\limits_{i=1}^m \frac{Np_{i0}^a}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 \xrightarrow{P} 0$ *and*
- *the distribution of (4.3.1) is estimated to be approximately $c\mathcal{X}_{m-1}^2$, where*

$$c = \frac{\min\limits_i p_{i0}^a + \max\limits_i p_{i0}^a}{2},$$

$\mathcal{X}_{m-1}^2$ is the chi-square distribution with $m - 1$ degrees of freedom and $\prec_{\text{st}}$ the symbol for stochastic ordering.

Proof. The Taylor expansion of Φ in an open ball $B_\varepsilon(p_i/p_{i0})$ of radius ε around the point $p_i/p_{i0},\ i = 1, 2, ..., m$, is given by:

$$\begin{aligned}\Phi\left(\frac{\hat{p}_i}{p_{i0}}\right) =& \Phi\left(\frac{p_i}{p_{i0}}\right) + \left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)\Phi'\left(\frac{p_i}{p_{i0}}\right) \\ &+ \frac{1}{2}\left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)^2 \Phi''\left(\frac{p_i}{p_{i0}}\right) + o\left(\left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)\right)^2.\end{aligned}$$

Multiplying both sides of the above relation by Np_{i0}^{1+a}, and taking the sum of both sides for $i = 1, 2, ..., m$ we get

$$\begin{aligned}\sum_{i=1}^{m} Np_{i0}^{1+a}\Phi\left(\frac{\hat{p}_i}{p_{i0}}\right) =& \sum_{i=1}^{m} Np_{i0}^{1+a}\Phi\left(\frac{p_i}{p_{i0}}\right) + \sum_{i=1}^{m} Np_{i0}^{1+a}\left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)\Phi'\left(\frac{p_i}{p_{i0}}\right) \\ &+ \frac{1}{2}\sum_{i=1}^{m} Np_{i0}^{1+a}\left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)^2\Phi''\left(\frac{p_i}{p_{i0}}\right) \\ &+ \sum_{i=1}^{m} Np_{i0}^{1+a} o\left(\left(\frac{\hat{p}_i}{p_{i0}} - \frac{p_i}{p_{i0}}\right)\right)^2.\end{aligned}$$

which for $p_i = p_{i0}$ becomes:

$$\begin{aligned}N\hat{d}_a - N\Phi(1)\sum_{i=1}^{m} p_{i0}^{1+a} - \frac{1}{2}\Phi''(1)\sum_{i=1}^{m}\frac{Np_{i0}^a}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2 \\ = N\sum_{i=1}^{m} p_{i0}^a\left(\frac{n_i}{N} - p_{i0}\right)\Phi'(1) + \sum_{i=1}^{m} N\frac{p_{i0}^a}{p_{i0}} o((\hat{p}_i - p_{i0}))^2.\end{aligned} \tag{4.3.2}$$

where $\hat{p} = (n_1/N, \ldots, n_m/N)'$ and $p_0 = (p_{10}, \ldots, p_{m0})'$. But

$$\begin{aligned}\sum_{i=1}^{m} N\frac{p_{i0}^a}{p_{i0}} o((\hat{p}_i - p_{i0}))^2 &\leqslant \max_i\left\{\frac{p_{i0}^a}{p_{i0}}\right\}\sum_{i=1}^{m} No((\hat{p}_i - p_{i0}))^2 \\ &= \max_i\left\{\frac{p_{i0}^a}{p_{i0}}\right\}\cdot N\cdot o(||\hat{p} - p_0||)^2 = o_P(1)\end{aligned} \tag{4.3.3}$$

since

$$\sqrt{N}(\hat{p} - p_0) \xrightarrow{L} N(0, \Sigma)$$

where Σ as in the proof of Theorem 4.3.1 (see Serfling, 1980, p. 108-109). From (4.3.2) and (4.3.3) we conclude that

$$\frac{2N\left(\hat{d}_a - \Phi(1)\sum\limits_{i=1}^{m} p_{i0}^{1+a}\right)}{\Phi''(1)} - \sum_{i=1}^{m}\frac{Np_{i0}^a}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2 \xrightarrow{P} 0.$$

Observe that

$$\left(\min_i p_{i0}^a\right)\sum_{i=1}^{m}\frac{N}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2 \prec_{\text{st}} \sum_{i=1}^{m}\frac{Np_{i0}^a}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2 \prec_{\text{st}} \left(\max_i p_{i0}^a\right)\sum_{i=1}^{m}\frac{N}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2.$$

The estimation of the distribution follow from the fact that as $N \to \infty$ (Serfling, 1980, page 122, example B)

$$\sum_{i=1}^{m}\frac{N}{p_{i0}}\left(\frac{n_i}{N} - p_{i0}\right)^2 \xrightarrow{L} \mathcal{X}_{m-1}^2.$$

■

Observe that in the theorem above we assume that $\Phi'(1) = 0$. This assumption is necessary if the test statistic used is the one given by (4.3.1). It is easy to see and it will be evident immediately after the Theorem 4.3.4 that this assumption is satisfied not only for the discrete BHHJ measure but also for all measures covered by the Csiszár's family of measures. If though one selects a function Φ which does not satisfy this assumption then the appropriate test statistic has to be defined. It is not difficult to see that in such a case (4.3.2) is the main expression affected since the first term on the right hand side of the expression does not vanish. The resulting test statistic will be given by

$$\Psi_a^2 \equiv \frac{2N\left(\hat{d}_a - \Phi\left(1\right)\sum\limits_{i=1}^{m} p_{i0}^{1+a} - \sum_{i=1}^{m} p_{i0}^{a}\left(\frac{n_i}{N} - p_{i0}\right)\Phi'(1)\right)}{\Phi''\left(1\right)}. \qquad (4.3.4)$$

It should be noted though that for values of a close to zero the last term in the numerator of (4.3.4) vanishes since

$$\sum_{i=1}^{m} p_{i0}^{a}\left(\frac{n_i}{N} - p_{i0}\right) \approx 0.$$

Theorem 4.3.3. *The power of the test*

$$H_0 : p_i = p_{i0} \quad vs \quad H_a : p_i = p_{ib}, \; i = 1, ..., m$$

using the test statistic (4.3.1) is approximately equal to:

$$\gamma_a = P\left(Z \geq \frac{\Phi''\left(1\right)c\mathcal{X}_{m-1,\alpha}^{2} + 2N\Phi\left(1\right)\sum\limits_{i=1}^{m} p_{i0}^{1+a} - 2Nd_a}{2\sqrt{N}\sigma_a}\right) \qquad (4.3.5)$$

where Z a standard Normal random variable, $\mathcal{X}_{m-1,\alpha}$ the $(1-\alpha)-$percentile of the $\mathcal{X}_{m-1}^{2}$ distribution, and

$$\sigma_a^2 = \sum_{i=1}^{m} p_{ib}\left[p_{i0}^{a}\Phi'\left(\frac{p_{ib}}{p_{i0}}\right)\right]^2 - \left[\sum_{i=1}^{m} p_{ib}p_{i0}^{a}\Phi'\left(\frac{p_{ib}}{p_{i0}}\right)\right]^2.$$

Proof. By definition, the power is given by

$$\begin{aligned}\gamma_a &= P\left(\mathrm{X}_a^2 \geq c\mathcal{X}_{m-1,\alpha}^{2}\middle| p_i = p_{ib}, i = 1, ..., m\right)\\ &= P\left(\hat{d}_a \geq \frac{\Phi''\left(1\right)c\mathcal{X}_{m-1,\alpha}^{2} + 2N\Phi\left(1\right)\sum\limits_{i=1}^{m} p_{i0}^{1+a}}{2N}\middle| p_i = p_{ib}, i = 1, ..., m\right).\end{aligned}$$

From Corollary 4.3.1 with $p_j = p_{jb},\ j = 1, ..., m$, we have

$$\frac{\sqrt{N}\left[\hat{d}_a - d_a\right]}{\sigma_a} \xrightarrow{L} N(0,1).$$

The result is immediate. ■

Note that for the BHHJ test corresponding to the measure given in (1.5.6) and (1.5.4) we have

$$\Phi''(1) = 1 + a \text{ and } \Phi(1) = \Phi'(1) = 0$$

so that the BHHJ statistic corresponding to the goodness of fit test of Theorem 4.3.2 is given by

$$\mathrm{X}_a^2 \equiv \frac{2N\hat{d}_a}{1+a} \tag{4.3.6}$$

while its power is given by

$$\gamma_a = P\left(Z \geq \frac{(1+a)c\mathcal{X}^2_{m-1,\alpha} - 2Nd_a}{2\sqrt{N}\sigma_a}\right). \tag{4.3.7}$$

Note also that the Csiszár's statistic corresponding to the goodness of fit test of Theorem 4.3.2 is given by

$$\mathrm{X}_c^2 \equiv \frac{2N\left(\hat{d}_c - \varphi(1)\right)}{\varphi''(1)} \tag{4.3.8}$$

while its power is given by

$$\gamma_c = P\left(Z \geq \frac{\varphi''(1)\mathcal{X}^2_{m-1,\alpha} + 2N\varphi(1) - 2Nd_c}{2\sqrt{N}\sigma_a}\right), \tag{4.3.9}$$

where

$$d_c = \sum_{i=1}^{m} p_{i0}\varphi(p_i/p_{i0}) \text{ and } \hat{d}_c = \sum_{i=1}^{m} p_{i0}\varphi(\hat{p}_i/p_{i0}).$$

For the usual Kullback-Leibler, Kagan and Cressie and Read measures we can easily see that

$$\varphi(1) = 0 \text{ and } \varphi''(1) = 1$$

so that the power is simplified into the form

$$\gamma_c = P\left(Z \geq \frac{\mathcal{X}^2_{m-1,\alpha} - 2Nd_c}{2\sqrt{N}\sigma}\right) \tag{4.3.10}$$

where

$$\sigma^2 = \sum_{i=1}^{m} p_{ib}\left[\varphi'\left(\frac{p_{ib}}{p_{i0}}\right)\right]^2 - \left[\sum_{j=1}^{m} p_{ib}\varphi'\left(\frac{p_{ib}}{p_{i0}}\right)\right]^2$$

and

$$\varphi'(x) = \log x \ (Kullback - Leibler),$$

$$\varphi'(x) = x - 1 \ (Kagan),$$

$$\varphi'(x) = \frac{1}{\lambda}(x^{\lambda} - 1) \ (Cressie\ and\ Read).$$

For the square of the Matusita measure it is not difficult to provide the appropriate expressions for the test statistic and the power since we can easily see that

$$\varphi(1) = 0,\ \varphi'(x) = 1 - x^{-1/2} \text{ and } \varphi''(1) = \frac{1}{2}.$$

We turn now to a special type of alternative hypothesis for multinomial populations. Suppose that the null hypothesis indicates that $p_i = p_{i0}$, $i = 1, 2, \ldots, m$ when in fact it is $p_i = p_{in}$, $\forall i$. As it is well known if p_{i0} and p_{in} are fixed then as n tends to infinity then the power of the test tends to 1. In order to examine the situation when the power is not close to 1, we must make it continually harder for the test as n increases. This can be done by allowing the alternative hypothesis steadily closer to the null hypothesis. As a result we define a sequence of alternative hypotheses as follows

$$H_{1,n} : p_i = p_{in} = p_{i0} + d_i/\sqrt{n},\ \forall i \qquad (4.3.11)$$

which is known as Pitman transition alternative or Pitman (local) alternative or local contiguous alternative to the null hypothesis $H_0 : p_i = p_{i0}$. In vector notation the local contiguous alternative takes the form

$$H_{1,n} : p = p_n = p_0 + d/\sqrt{n}$$

and the null the form

$$H_0 : p = p_0$$

where $p = (p_1, \ldots, p_m)'$, $p_n = (p_{1n}, p_{2n}, \ldots, p_{mn})'$, and $d = (d_1, \ldots, d_m)'$ is a fixed vector such that $\sum_{i=1}^{m} d_i = 0$. Observe that as n tends to infinity the local contiguous alternative converges to the null hypothesis at the rate $O(n^{-1/2})$.

We define now the noncentral chi-square distribution.

Definition 4.3.3. *If $X_1, \ldots, X_m$ are independent random variables with $X_i \sim N(\xi_i, 1)$, the distribution of $\sum_{i=1}^{m} X_i^2$ is noncentral chi-square with m degrees of freedom and noncentrality parameter $\delta = \sum_{i=1}^{m} \xi_i^2$. In matrix notation we say that if $X \sim N(\xi, I)$ then $X'X \sim \mathcal{X}_{m,\delta}^2$, with $\delta = \xi'\xi$ where $X = (X_1, \ldots, X_m)'$, $\xi = (\xi_1, \ldots, \xi_m)'$ and I the mxm identity matrix.*

The following Lemma from Hunter (2002, p. 72) which will be used later is presented below without proof. The lemma provides conditions for the noncentral chi-square distribution but applies also to the chi-square distribution when ξ is taken to be 0.

Lemma 4.3.1. *Suppose that $X \sim N(\xi, Q)$ where Q is a projection matrix of rank $r \leq m$ and $Q\xi = \xi$. Then, $X'X \sim \mathcal{X}_{r,\xi'\xi}^2$.*

In order to derive the asymptotic distribution of the test statistic (4.3.1) under the local contiguous alternatives $H_{i,n}$, observe that when indeed $p_i = p_{in}$, $\forall i$ and $\hat{p}_i$ the maximum likelihood estimator of p_i then

$$\sqrt{n}\frac{(\hat{p}_i - p_{in})}{\sqrt{p_{in}(1 - p_{in})}} \xrightarrow{L} N(0, 1).$$

Observe also that

$$\sqrt{\frac{p_{in}}{p_{i0}}} = \sqrt{1 + \frac{p_{in} - p_{i0}}{p_{i0}}} = \sqrt{1 + \frac{d_i}{\sqrt{n}p_{i0}}}$$

which converges to 1 as $n \to \infty$. In a similar fashion one can easily show that

$$\sqrt{\frac{1 - p_{in}}{1 - p_{i0}}} = \sqrt{1 - \frac{d_i}{\sqrt{n}(1 - p_{i0})}}$$

which converges also to 1 as $n \to \infty$. As a result

$$\sqrt{n}\frac{(\hat{p}_i - p_{in})}{\sqrt{p_{in}(1 - p_{in})}} \cdot \frac{\sqrt{p_{in}(1 - p_{in})}}{\sqrt{p_{i0}(1 - p_{i0})}} \xrightarrow{L} N(0, 1)$$

or equivalently

$$\sqrt{n}\frac{(\hat{p}_i - p_{in})}{\sqrt{p_{i0}(1 - p_{i0})}} \xrightarrow{L} N(0, 1).$$

It is easily seen that

$$\begin{aligned}\sqrt{n}(\hat{p}_i - p_{i0}) &= \sqrt{n}(\hat{p}_i - p_{in}) + \sqrt{n}(p_{in} - p_{i0}) \\ &= \sqrt{n}(\hat{p}_i - p_{in}) + d_i.\end{aligned}$$

Hence, by Slutsky's theorem

$$\sqrt{n}(\hat{p}_i - p_{i0}) \xrightarrow{L} N(d_i, p_{i0}(1-p_{i0})).$$

Furthermore, observe that $Cov\,(\hat{p}_i - p_{i0})\,(\hat{p}_j - p_{j0}) = n^{-1}p_{i0}p_{j0}$. In conclusion for the m-dimensional vector parameter we have (see also Serfling, 1980, pp. 108-109)

$$\sqrt{n}(\hat{p} - p_n) \xrightarrow{L} N(0, \Sigma)$$

and

$$\sqrt{n}(\hat{p} - p_0) \xrightarrow{L} N(d, \Sigma)$$

where $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)'$ and Σ as in the proof of Theorem 4.3.1. Let P a diagonal matrix with diagonal elements the inverses of the elements of the vector p_0. Then, from Theorem 4.3.2 we have

$$\begin{aligned} \sum_{i=1}^{m} \frac{N}{p_{i0}} \left(\frac{n_i}{N} - p_{i0}\right)^2 &= N(\hat{p} - p_0)' P (\hat{p} - p_0) \\ &= \left(\sqrt{N}\left(P^{1/2}(\hat{p} - p_0)\right)'\right)\left(\sqrt{N}\left(P^{1/2}(\hat{p} - p_0)\right)\right) \end{aligned}$$

so that by Slutsky's theorem

$$\sqrt{N}\left(P^{1/2}(\hat{p} - p_0)\right) \xrightarrow{L} N(P^{1/2}d, P^{1/2}\Sigma P^{1/2}).$$

Lemma 4.3.1 can now be applied provided that the matrix $P^{1/2}\Sigma P^{1/2}$ is of rank $m-1$ and that $(P^{1/2}\Sigma P^{1/2}) \cdot (P^{1/2}d) = P^{1/2}d$.

For the first condition we have

$$P^{1/2}\Sigma P^{1/2} = P^{1/2}[P^{-1} - p_0 p_0']P^{1/2} = I - P^{1/2}p_0 p_0' P^{1/2} = I - \sqrt{p_0}\sqrt{p_0}'$$

which clearly is symmetric with trace equal to $m-1$. The sum of its eigenvalues is also equal to $m-1$ since for symmetric matrices the trace and the sum of the eigenvalues coincide. Furthermore, since $\sqrt{p_0}'\sqrt{p_0} = 1$ we have that

$$(I - \sqrt{p_0}\sqrt{p_0}')(I - \sqrt{p_0}\sqrt{p_0}') = I - 2\sqrt{p_0}\sqrt{p_0}' + \sqrt{p_0}\sqrt{p_0}'\sqrt{p_0}\sqrt{p_0}' = I - \sqrt{p_0}\sqrt{p_0}'$$

and hence, the matrix $P^{1/2}\Sigma P^{1/2}$ is a projection matrix with implies that its eigenvalues are all equal to 0 or 1. As a result there are $m-1$ eigenvalues equal to 1.

The second condition is easily established since

$$P^{1/2}\Sigma P d = P^{1/2}[P^{-1} - p_0 p_0']Pd = P^{1/2}[d - p_0(1)'d]$$

where the second term vanishes since $(1)'d = \sum_{i=1}^{m} d_i = 0$, $\Sigma = P^{-1} - p_0 p_0'$ the covariance matrix appearing in the proof of the Theorem 4.3.1 and (1) an m-dimensional vector with elements equal to 1.

As a result, in contrast to the chi-square distribution derived in Theorem 4.3.2, here and as $N \to \infty$ and under the local contiguous alternative hypotheses $H_{i,n}$ we observe the non-central distribution, namely

$$\sum_{i=1}^{m} \frac{N}{p_{i0}} \left(\frac{n_i}{N} - p_{i0} \right)^2 \xrightarrow{L} \mathcal{X}^2_{m-1,\delta},$$

where the noncentrality parameter δ is given by

$$\begin{aligned} \delta = (P^{1/2}d)'P^{1/2}d &= d'Pd \\ &= \sum_{i=1}^{m} \frac{d_i^2}{p_{i0}}. \end{aligned}$$

The following theorem summarizes the above discussion:

Theorem 4.3.4. *The asymptotic distribution of the test statistic given in (4.3.1) under the local contiguous alternative hypotheses (4.3.11), is* $c\mathcal{X}^2_{m-1,\delta}$ *where* $c = \frac{\min_i p^a_{i0} + \max_i p^a_{i0}}{2}$ *and* $\mathcal{X}^2_{m-1,\delta}$ *is the noncentral chi-square distribution with* $m-1$ *degrees of freedom and noncentrality parameter given by* $\delta = \sum_{i=1}^{m} \frac{d_i^2}{p_{i0}}$.

Following the above theorem the power of the test under the local contiguous alternative hypotheses (4.3.11) is given by

$$\begin{aligned} \gamma_n = P(\mathrm{X}_a^2 > c\mathcal{X}^2_{m-1,\alpha} | p_i = p_{in}, i = 1, ..., m) &= P(c\mathcal{X}^2_{m-1,\delta} > c\mathcal{X}^2_{m-1,\alpha}) \\ &= P(\mathcal{X}^2_{m-1,\delta} > \mathcal{X}^2_{m-1,\alpha}). \quad (4.3.12) \end{aligned}$$

Note that the corresponding power of the above test using the Csiszár's statistic (4.3.8) is given by exactly the same formula, namely

$$\gamma_n = P(\mathrm{X}_c^2 > \mathcal{X}^2_{m-1,\alpha} | p_i = p_{in}, i = 1, ..., m) = P(\mathcal{X}^2_{m-1,\delta} > \mathcal{X}^2_{m-1,\alpha}). \quad (4.3.13)$$

Chapter 5

Simulations

5.1 Model Selection

In order to check the performance of the DIC criterion proposed in Section 3.2 we performed a simulation study using

- the Divergence Information Criterion DIC
- The corrected DIC_c based on the MLE method
- the Akaike Information Criterion AIC
- the Bayesian Information Criterion BIC
- the AIC for small sample sizes and
- the AIC with the estimator of the variance obtained by the minimization of the BHHJ measure.

The simulation study has the following characteristics. 50 observations of 4 variables X_1, X_2, X_3, X_4 were independently generated from the normal distributions $N(0,3)$, $N(1,3)$, $N(2,3)$ and $N(3,3)$ correspondingly. Correlation coefficients between these variables were less than 0.13 (in absolute values) in all cases. The first 2 of these variables was planned to be used to generate values of Y_i, $i = 1, \ldots, 50$ using the following model specification:

$$Y_i = a_0 + a_1 X_{1,i} + a_2 X_{2,i} + \varepsilon_i$$

with

$$a_0 = a_1 = a_2 = 1 \text{ and } \varepsilon_i \sim N(0,1).$$

Due though to contamination of the above model by 10% from the model

$$Y_i = 1 + X_{1,i} + X_{2,i} + \varepsilon_i^*$$

with $\varepsilon_i^* \sim N(5,1)$ the simulated values were generated from the model

$$Y_i = 0.9(1 + X_{1,i} + X_{2,i} + \varepsilon_i) + 0.1(1 + X_{1,i} + X_{2,i} + \varepsilon_i^*).$$

The reason for introducing contamination into the simulation study was to put into a test the robust features of the DIC criterion. In other words, we wanted to force the DIC to perform to the fullest extent and activate its prime feature according to which when $a > 0$, observations significantly discrepant with respect to the model get an almost zero weight and therefore their contribution to the final selection is minimal.

With a set of 4 possible regressors there are $2^4 - 1 = 15$ possible specifications that include at least one regressor. These 15 possible regression specifications constitute the set of candidate models for the experiment. As a result the candidate set consists of the full model (X_1, X_2, X_3, X_4) given by

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + \varepsilon$$

as well as all 14 possible subsets of the full model consisting of one (X_{j_1}), two (X_{j_1}, X_{j_2}) and three $(X_{j_1}, X_{j_2}, X_{j_3})$, with $j_i \in \{1,2,3,4\}, i = 1,2,3$ of the 4 regressors X_1, X_2, X_3, X_4. 50 such experiments were performed with the intention to select the best model among the available candidate models.

Recall that the construction of DIC is similar in spirit to the construction of AIC since they are both established by obtaining an unbiased estimator of the overall discrepancy. Furthermore, the consistency property of BIC makes it a highly applicable criterion. As a result it is highly desirable to compare the three criteria in terms of their performance. Besides the standard AIC criterion two more variations of AIC have also been included in the analysis.

First we consider the standard AIC criterion given by

$$AIC = n \log \hat{\sigma}_p^2 + 2\,(p + 2)$$

where n the sample size, p the number of variables of the model and $\hat{\sigma}_p^2$ the estimate of the variance of the model with p variables.

We also consider the corrected AIC criterion introduced by Hurvich and Tsai (1989) and used in small sample situations. The corrected AIC is given by

$$AIC_c = n\log\hat{\sigma}_p^2 + \frac{n(n+p+1)}{n-p-3}.$$

Another variant of the AIC criterion used in the simulations is the one given by

$$AIC_a = n\log\hat{\sigma}_{p,a}^2 + 2\,(p+2)$$

where $\hat{\sigma}_{p,a}^2$ is the estimator of the variance $\sigma_{p,a}^2$ of the model with p variables which is obtained by the minimization of the BHHJ measure. Note that there is no closed form for the estimators of the parameters but they are computed by using numerical methods to solve the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} u_{\boldsymbol{\theta}}\,(X_i) f_{\boldsymbol{\theta}}^{a}\,(X_i) - \int u_{\boldsymbol{\theta}}\,(z) f_{\boldsymbol{\theta}}^{1+a}\,(z)\,dz = 0$$

where $u_{\boldsymbol{\theta}}\,(z) = \frac{\partial\log f_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}}$ and $\boldsymbol{\theta} = (b_0,\ldots,b_p,\sigma_{p,a}^2)$, $p = 1,2,3,4$. AIC_a is evaluated for $a = 0.01, 0.05$ and 0.10.

From the various Bayesian approaches we have chosen to include in the simulations the Bayesian Information Criterion (BIC, Schwarz, 1978) because of its consistency property. The BIC is given by

$$BIC = n\log\hat{\sigma_p}^2 + (p+2)\log n.$$

Finally the DIC is used with both corrected and uncorrected penalty terms and with both estimating methods, namely the Basu and the MLE methods. The original DIC (uncorrected) based on the Basu method, is used with index $a = 0.01, 0.05$ and 0.10 and the corrected DIC_c based on the MLE method, with $a = 0.01, 0.05, 0.10$ and 0.15. To make the notation precise we will be using in the sequel DIC_c^{MLE} in place of DIC_c. Recall that the formulas of the DIC criterion are given by

$$DIC = nQ_{\hat{\theta}} + (a+1)\,(2\pi)^{-\frac{a}{2}}\left(\frac{1+a}{1+2a}\right)^{1+\frac{p}{2}} p$$

and

$$DIC_c^{MLE} = nQ_{\hat{\theta}} + (2\pi)^{-\frac{a}{2}}\,(1+a)^{2+\frac{p}{2}}\,p.$$

For each of the 50 experiments the value of each of the above model selection criteria was calculated for each of the 15 possible regression specifications under consideration. As a result, for each of the 50 experiments and for each model selection criterion the 15 candidate models were ranked from 1st to 15th according to the value of the criterion. Recall that the model chosen by a criterion is the one for which the value of the criterion is the lowest among all 15 candidate models. Table 5.1 presents for each selection criterion, the proportion of times each candidate model has been selected by the criterion. Notice that only 4 of the 15 candidate models have been ranked 1st and therefore selected, namely the true model (X_1, X_2), and the "larger" models (X_1, X_2, X_3), (X_1, X_2, X_4) and (X_1, X_2, X_3, X_4). Obviously, all selections contain the correct variables of the model, namely X_1 and X_2.

Observe that the DIC criterion selects the true model in all instances where the AIC criterion succeeds, that is 80% of the cases. The AIC_c has a higher success rate (88%) which could be attributed to the relative small sample size used (n=50). The AIC criterion with index a has the smaller rate of success (less than 80%). In fact observe that the larger the value of the index a the worse the performance of the resulting criterion.

On the other hand both BIC and DIC_c^{MLE} with $a = 0.15$ have the best selection rate (96%) among all competing selection criteria. It should be noted that for DIC the selection rate improves as a tends to 0 while for DIC_c^{MLE} the rate improves as a increases up to a maximum value. This behavior is due to the different form of the correction term. Indeed, DIC decreases as a function of the index a while DIC_c^{MLE} is an increasing function of a. As a result and as a (and p) increases, the DIC_c^{MLE} criterion puts a heavier penalty in large models (in models where the dimension p of the parameter is large) and therefore for too large values of a (and p) we end up underestimating the true model.

The performance of DIC_c^{MLE} seems to be superior than that of DIC not only because of its higher rate of success but also because it is based on the MLE method which is computationally faster than the Basu method since the former is provided in closed form while the latter relies on a numerical method for obtaining the required estimator.

In conclusion, the DIC expresses a good medium sample size performance which is comparable to the traditional AIC criterion while the DIC_c^{MLE} is very powerful and comparable to BIC.

Table 5.1: Proportion of the selected models by model selection criteria (n=50)

Criteria		Variables	%	Variables	%	Variables	%
AIC	AIC	X_1, X_2	80	X_1, X_2, X_4	20	*	*
	AIC_c	X_1, X_2	88	X_1, X_2, X_4	12	*	*
BIC	BIC	X_1, X_2	96	X_1, X_2, X_4	4	*	*
AIC_a	$AIC_{0.01}$	X_1, X_2	80	X_1, X_2, X_4	16	X_1, X_2, X_3	4
	$AIC_{0.05}$	X_1, X_2	76	X_1, X_2, X_4	16	X_1, X_2, X_3	8
	$AIC_{0.10}$	X_1, X_2	68	X_1, X_2, X_4	16	X_1, X_2, X_3 or X_1, X_2, X_3, X_4	16
DIC	$DIC_{0.01}$	X_1, X_2	80	$X_1,\ X_2,\ X_4$	20	*	*
	$DIC_{0.05}$	X_1, X_2	76	$X_1,\ X_2,\ X_4$	20	X_1, X_2, X_3	4
	$DIC_{0.10}$	X_1, X_2	72	$X_1,\ X_2,\ X_4$	16	X_1, X_2, X_3 or X_1, X_2, X_3, X_4	12
DIC_c^{MLE}	$DIC_{0.01}$	X_1, X_2	80	$X_1,\ X_2,\ X_4$	20	*	*
	$DIC_{0.05}$	X_1, X_2	80	$X_1,\ X_2,\ X_4$	20	*	*
	$DIC_{0.10}$	X_1, X_2	88	$X_1,\ X_2,\ X_4$	12	*	*
	$DIC_{0.15}$	X_1, X_2	96	$X_1,\ X_2,\ X_4$	4	*	*

5.2 Goodness of Fit Tests

For checking the accuracy of the proposed BHHJ test of Section 4.3 theoretical and simulated results using trinomial distributions are obtained in the present section. In particular and in order to understand the behavior of the BHHJ test we compare it with four other tests, namely the goodness of fit tests based on

- the Kullback-Leibler measure (KL),
- the Kagan measure,
- the Matusita measure (Mat) and
- the Cressie and Read measure with $\lambda = 2/3$ (CR).

The proposed BHHJ goodness of fit test is applied for three different values of the index a, namely for $a = 0.01, 0.05$ and 0.10. Both the power and the type I error are investigated. For the theoretical (asymptotic) power formulas (4.3.7), (4.3.9) and (4.3.10) are used. For the simulated results for both the power and the type I error of the test the sample size from the trinomial distribution used is equal to 150 and a number of 10000 simulations have been created. The large number of simulations is explained by the fact that the theoretical power was required to be checked for accuracy. The following null hypothesis is assumed

$$H_0 : p_{10} = 0.2, p_{20} = 0.6, p_{30} = 0.2.$$

The various alternatives used are presented in Table 5.2 (p_{3b} is omitted since $\sum_{i=1}^{3} p_{ib} = 1$).

In Table 5.2 the theoretical powers of the above tests are presented along with the powers based on the simulated study. Table 5.2 provides also for comparative purposes the theoretical power calculated by equation (4.3.13) of the test under the local contiguous alternative hypotheses. This technique is used with $n = 150$ under the alternative hypothesis $H_1 : p_i = p_{ib}$, $i = 1, \ldots, m$ which is viewed as a contiguous alternative with

$$d = \sqrt{n}(p_{1b} - p_{10}, p_{2b} - p_{20}, p_{3b} - p_{30})'.$$

This test with the above alternative will be refer to in the sequel as *Test with Contiguous Alternatives* and denoted by *TCA* in Table 5.2.

Table 5.2 compares the central values of the BHHJ test with $a = 0.01, 0.05$ and 0.10 and the 4 competing tests while Table 5.4 provides the results for the BHHJ test for $a = 0.01$ (only the central BHHJ-C value is provided), $a = 0.05$ and $a = 0.10$.

Observe that in the case of the BHHJ statistic Table 5.4 provides the asymptotic power of the test (BHHJ-C) as well as the upper (BHHJ-U) and lower (BHHJ-L) limits of the asymptotic power as they can be deduced from Theorem 4.3.2.

The tables provide the probabilities for the former case while for the latter the tables provide the number of times (out of 10000) the null hypothesis is rejected.

The results from the power calculations reveal a number of conclusions which are stated below:

- It can be easily seen that the simulated results are much better that the theoretical ones for all 5 competing tests. This observation indicates that these power approximations all of which are based on the normal distribution, are not the best possible.

- The theoretical and simulated results for the Kagan test represent also the corresponding results for the Pearson's chi-square test since the two tests are identical. Recall that the theoretical power for all 4 competing tests in Table 5.2 were calculated using equations (4.3.7) and (4.3.9). The inclusion in our analysis of the Test with Contiguous Alternatives is due to our effort to compare the results, both theoretical and simulated, of the 4 competing tests to the theoretical results based on the theory of contiguous alternatives. Note that the equations for the evaluation of the power of the Test with Contiguous Alternatives given by (4.3.12) and (4.3.13) imply that the power using this technique is exactly the same irrespectively of the form of the function Φ and the value of the index α used in (4.3.1). It seems that in almost all cases the power theoretical results of the Test with Contiguous Alternatives are closer to the power simulated results obtained by each of the 4 competing tests.

- The BHHJ statistic performs in simulations better than the Kullback-Leibler statistic irrespectively of the alternative hypothesis. Note that this is also evident from the theoretical calculations.

- The BHHJ test performs better than all other tests for all alternatives that are not far away from the null hypothesis. On the other hand it performs as good as all other tests for all alternatives that are far away from the null hypothesis.

- Both the theoretical and the simulated results show that the Matusita and the BHHJ tests have a very similar behavior and in most cases are the most powerful tests among the ones examined. Both tests behave well for alternatives close to the null hypothesis and better than the Kullback-Leibler test. This observation indicates that the BHHJ test, as well as the Matusita test, is able to distinguish between null and alternative hypotheses when they are very close.

- Recall that for the BHHJ statistic the central as well as the upper (BHHJ-U) and lower (BHHJ-L) bounds of the power are provided for both the simulated and the theoretical calculations in Table 5.4. It should be noted that the lower bound depends on the smaller of the probabilities p_{i0} of the null multinomial distribution while the upper bound depends on the larger of these probabilities. Note though that for values of a close to zero (as it is the case in most applications) the corresponding quantities involved in the evaluation of the bounds, namely $\min p_{i0}^a$ and $\max p_{i0}^a$ are equal to a value not far from 1 and consequently the associated bounds are not far from the central value BHHJ-C. Finally note that besides the average value proposed in Theorem 4.3.2 we could easily use for the evaluation of the power of the BHHJ test, the middle value (the median value) of the probabilities of the multinomial distribution under the null hypothesis.

- Finally observe that the larger the value of a the smaller the power of the BHHJ test and the larger the range between the upper and the lower limits of the power of the BHHJ test.

Simulations have also been used to evaluate the type I error of the proposed BHHJ test. The results presented in Table 5.3 and in Table 5.5 for various null hypotheses in the case of the trinomial distribution show that all tests perform quite well with sizes around the typical 5% level. Observe that in this case the larger the value of the index a for the BHHJ test statistic the smaller the type I error (in the expense of smaller power).

Since all tests do not have the correct size, it is desirable to make the necessary power adjustment in order to compare properly the competing tests. One of the graphical methods used for comparing the power of competing tests is the so called size-power curve. These curves are constructed using empirical distribution functions (EDF), one for an experiment where the null is true and one for an experiment where the null is false. Let $F_1(x)$ and $F_2(x)$ the two EDFs evaluated at pre-chosen points $x_1, \ldots, x_n$. $F_1(x)$ is the probability of getting a P-value less than x under the null. Similarly, $F_2(x)$ is the corresponding probability under the alternative. Tracing the locus $(F_1(x), F_2(x))$ inside the unit square as x varies from 0 to 1 we generate the size-power curve with a correct size-adjusted basis. The purpose of including the same points $x_i, i = 1, \ldots, n$ is the reduction of the experimental error. Note that the method of size-power curve has been introduced by Wilk and Gnanadesikan (1968).

We have included in the manuscript only 2 power-size curves. Fig. 5.1 corresponds to an alternative neither very close to the null nor far away from the null. More specifically, the test corresponding to Fig. 5.1 is

$$H_0 : p_1 = 0.2, p_2 = 0.6, p_3 = 0.2$$

$$vs.$$

$$H_1 : p_1 = 0.2, p_2 = 0.7, p_3 = 0.1.$$

Fig. 5.2 corresponds to an alternative very close to the null hypothesis and is selected to show that there are cases where there is not clear advantage in using one instead of another test. The test corresponding to Fig. 5.2 is

$$H_0 : p_1 = 0.2, p_2 = 0.6, p_3 = 0.2$$

$$vs.$$

$$H_1 : p_1 = 0.25, p_2 = 0.60, p_3 = 0.15.$$

The results clearly indicate that when the power is adjusted for size then the proposed BHHJ test retains its superiority being the most powerful among the 6 competing tests although there are cases (very close to the null hypothesis) where all tests perform equally well. It is interesting to note that Matusita's statistic performs quite well coming second after the BHHJ statistic. On the other hand though the Kagan's test (i.e. the Pearson's X^2 test) and the Cressie and Read test for $\lambda = 2/3$ have the worst performance among the 6 competing tests which puts into question the hypothesized superiority of these two tests (Cressie and Read, 1988, Chapters 5 and 6). Further investigation is required though to verify such a claim.

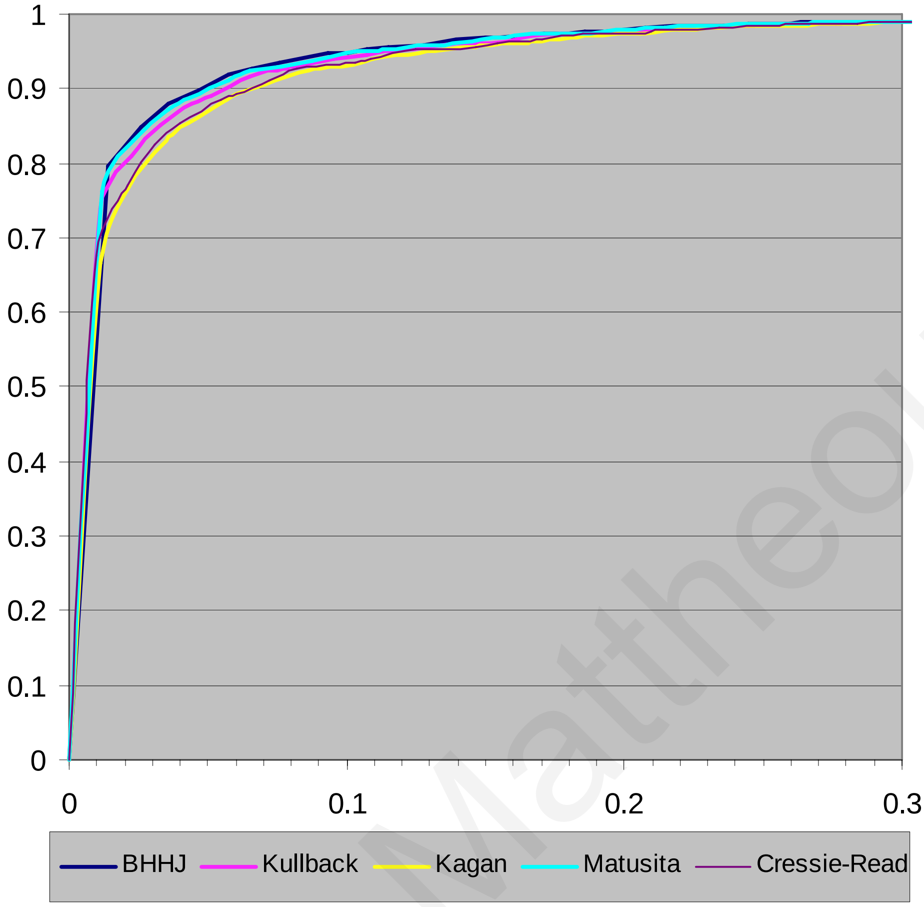

Figure 5.1: Power vs. Size curves for the comparison of competing tests (alternative not far away).

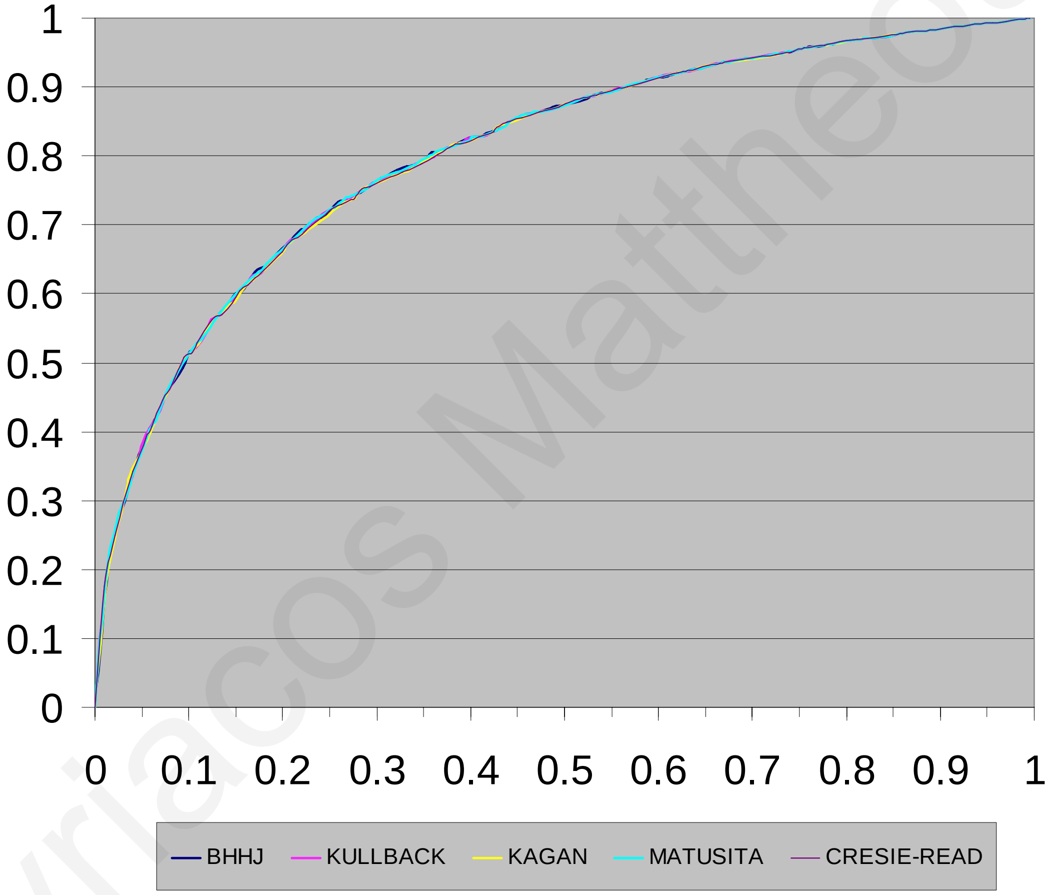

Figure 5.2: Power vs. Size curves for the comparison of competing tests (alternative close to the null).

Table 5.2: Theoretical (asymptotic) and simulated power calculations for trinomial distributions.

Alternative	Competing Tests					BHHJ Test		
$H_0 : p_{10} = 0.20, p_{20} = 0.60, p_{30} = 0.20,\ \mathcal{X}^2_{2;0.05},\ n = 150$								
p_{1b} & p_{2b}	KL	Kagan	Mat	CR	TCA	$a = 0.01$	$a = 0.05$	$a = 0.10$
0.21, 0.59	0	0	0	0	0.0576	0	0	0
0.22, 0.60	0.0003	0.0002	0.0003	0.0002	0.0980	0.0003	0.0002	0.0001
0.25, 0.60	0.284	0.279	0.290	0.279	0.392	0.295	0.282	0.265
0.20, 0.70	0.818	0.793	0.825	0.803	0.815	0.828	0.825	0.821
0.10, 0.60	0.901	0.890	0.900	0.895	0.944	0.896	0.894	0.891
0.40, 0.36	0.997	0.996	0.997	0.996	1	0.996	0.996	0.996
0.45, 0.35	0.999	0.999	0.999	0.999	1	0.999	0.999	0.999
0.40, 0.30	0.999	0.999	0.999	1	1	0.999	0.999	0.999
0.55, 0.25	1	1	1	1	1	1	1	1
SIMULATIONS (# of rejections of H_0 in 10000 samples)								
0.21, 0.59	607	686	649	628	*	629	595	573
0.22, 0.60	944	1000	1013	969	*	1009	968	913
0.25, 0.60	3880	3926	4027	3906	*	4042	3867	3746
0.20, 0.70	8915	8839	9110	8853	*	9187	9182	9181
0.10, 0.60	9638	9608	9665	9622	*	9665	9617	9596
0.40, 0.36	9999	10000	9999	10000	*	9999	9999	9999
0.45, 0.35	10000	10000	10000	10000	*	10000	10000	10000
0.40, 0.30	10000	10000	10000	10000	*	10000	10000	10000
0.55, 0.25	10000	10000	10000	10000	*	10000	10000	10000

Table 5.3: Type I error calculations for trinomial distributions.

Null	Competing Tests				BHHJ Test		
SIMULATIONS (# of rejections of H_0 in 10000 samples), $\mathcal{X}^2_{2;0.05}$, $n = 150$							
p_{10} & p_{20}	KL	Kagan	Mat	CR	$a = 0.01$	$a = 0.05$	$a = 0.10$
0.20, 0.60	512	567	578	544	575	558	524
0.30, 0.40	535	536	564	536	579	579	573
0.10, 0.80	499	512	569	453	589	555	476
0.10, 0.50	507	529	558	491	569	578	580
0.30, 0.35	518	504	538	513	560	560	558

Table 5.4: Theoretical (asymptotic) and simulated power calculations for trinomial distributions for the BHHJ test.

Alternative	$a = 0.01$	$a = 0.05$			$a = 0.10$		
$H_0 : p_{10} = 0.20, p_{20} = 0.60, p_{30} = 0.20$, $\mathcal{X}^2_{2;0.05}$, $n = 150$							
p_{1b} & p_{2b}	BHHJ-C	BHHJ-L	BHHJ-C	BHHJ-U	BHHJ-L	BHHJ-C	BHHJ-U
0.21, 0.59	0	0	0	0	0	0	0
0.22, 0.60	0.0003	0.0001	0.0002	0.0003	0.0001	0.0001	0.0003
0.25, 0.60	0.295	0.268	0.282	0.296	0.237	0.265	0.295
0.20, 0.70	0.828	0.819	0.825	0.830	0.809	0.821	0.832
0.10, 0.60	0.896	0.890	0.894	0.897	0.883	0.891	0.898
0.40, 0.36	0.996	0.996	0.996	0.996	0.996	0.996	0.996
0.45, 0.35	0.999	0.999	0.999	0.999	0.999	0.999	0.999
0.40, 0.30	0.999	0.999	0.999	0.999	0.999	0.999	0.999
0.55, 0.25	1	1	1	1	1	1	1
SIMULATIONS (# of rejections of H_0 in 10000 samples)							
0.21, 0.59	629	570	595	668	486	573	680
0.22, 0.60	1009	936	968	1050	787	913	1081
0.25, 0.60	4041	3819	3866	4083	3450	3745	4156
0.20, 0.70	9186	9147	9181	9203	8888	9080	9239
0.10, 0.60	9664	9606	9616	9670	9509	9595	9686
0.40, 0.36	9998	9998	9998	9998	9998	9998	9998
0.45, 0.35	9999	9999	9999	9999	9999	9999	9999
0.40, 0.30	9999	9999	9999	9999	9999	9999	9999
0.55, 0.25	9999	9999	9999	9999	9999	9999	9999

Table 5.5: Type I error calculations for trinomial distributions for the BHHJ test.

Null	$a = 0.01$	$a = 0.05$			$a = 0.10$		
SIMULATIONS (# of rejections of H_0 in 10000 samples), $\mathcal{X}^2_{2;0.05}$, $n = 150$							
p_{10} & p_{20}	BHHJ-C	BHHJ-L	BHHJ-C	BHHJ-U	BHHJ-L	BHHJ-C	BHHJ-U
0.20, 0.60	575	531	558	601	430	524	625
0.30, 0.40	579	573	579	589	548	573	589
0.10, 0.80	589	482	555	600	333	476	612
0.10, 0.50	569	520	578	602	456	580	706
0.30, 0.35	560	550	560	566	550	558	566

Discussion

The main topic of this thesis is the investigation of the BHHJ measure of divergence. We propose a general BHHJ family of measures of divergence that includes the BHHJ measure of divergence (Basu et. al, 1998) as well as the Csiszar's family of measures (Csiszar, 1963; Ali and Silvey, 1966). Furthermore, we propose a class of discrete measures of divergence which could be considered as the discrete version of the above mentioned general class of measures.

A number of properties of the general BHHJ class of measures has been discussed like the symmetry property, the limiting property, and the quadratic convergence. Furthermore, we propose a new model selection criterion called Divergence Information Criterion (DIC) which is based on the BHHJ measure. Finally we propose a test statistic for goodness of fit tests for multinomial populations. Note that both the DIC criterion and the test statistic are indexed by a single parameter a. The value of a dictates to what extent the estimating method based on the minimization of the measure of divergence becomes more robust than the maximum likelihood estimating method. One should be aware of the fact that the larger the value of a the bigger the efficiency loss. Consequently, one should be interested in small values of $a > 0$, say between zero and one.

The proposed DIC criterion could be used in applications where outliers or contaminated observations are involved. The prior knowledge of contamination may be useful in identifying an appropriate value of a. Simulations show that values of a from 0.01 to 0.10 are sufficient in achieving high success rate of correct model selections.

The proposed BHHJ test statistic was compared with other tests like the Pearson's X^2 test or Kagan's test, the loglikelihood ratio test (Kullback-Leibler test), the Cressie and Read test and the Matusita test and was found to perform well in cases where the alternative hypothesis is close or not far away from the null. In cases where the two hypotheses differ significantly, all tests, including the BHHJ test, perform equally well. Simulations based on trinomial distributions show that the proposed BHHJ test statistic is superior to other traditional goodness of fit tests when the power is adjusted for the size of the test.

Future Research

The results obtained in this thesis can be extended and generalized in a number of ways.

Regarding the DIC model selection criterion proposed in Chapter 3 one should investigate its asymptotic properties. Two of the issues in model selection that are discussed in the literature are consistency and asymptotic efficiency. A natural requirement for a selection procedure is to choose the best possible model from a given family of models. Needless to say, the goodness depends on the objective of the analysis. Consistency is our main concern whenever we know the true model as correctly as possible. In other words, consistency is of great importance if the true model belongs to the family of models from which the selection is to be made. On the other hand, the asymptotic efficiency is associated with the predictive performance and requires the selection of a model which yields good predictions. For this objective it is natural to assume that the true model does not necessarily coincide with one of the models under consideration. It is important to point out that the two issues are not compatible. In particular, Shibata (1976) and Bhansali and Downham (1977) showed that AIC and its alike tend asymptotically to overfit the true order (overestimation) and therefore they are inconsistent. A recent paper by Wei (1992) investigates the distributional properties of a number of criteria and establishes the consistency of BIC. In the same paper, the author proposes the use of a new criterion that incorporates the Fisher's Information (FIC) and proves its consistency. Notice the different meaning and different usage of the terms "consistency" and "efficiency" in the order selection theory. For example, the notion of "inconsistency" (overestimation) of AIC could be viewed as equivalent to "superconsistency" in the traditional sense. As a result, caution is required whenever such issues are raised so that unnecessary misinterpretations would be avoided. As a result, if the true model is unknown the concept of consistency should not be included at the top of the scientist's list. The asymptotic efficiency is solely associated with prediction and if this is the purpose of the study, then a selection strategy carrying such a property should be used.

The notion of asymptotic efficiency which was introduced by Shibata (1980) is based on the selection of that model which leads to the smallest average mean squared error of prediction. The theory developed in recent years (e.g. Shibata, 1981, Bhansali, 1986, Hurvich & Tsai, 1989, Karagrigoriou, 1997) shows that the family of AIC-type criteria possesses such a property as opposed to the family of BIC-type criteria which have been found to be consistent but not asymptotically efficient. This latter property however, is useful in practice only in the case where the class of candidate models does include the correct model. Since however the true underlying model is unknown in practice the notion of asymptotic efficiency seems to be a more realistic property.

Shibata (1980) was the first to make the innovative assumption that the data-generating mechanisms belong to a class of linear models with infinitely many unknown parameters. As a result the concept of asymptotic efficiency is associated with a finite approximation of the truly infinite order of the model and as such it is not an estimating but rather an approximation problem. The scope is to obtain a good approximation to the underlying model which could be potentially useful for predictive purposes. In all cases where a predictor or a modelling assessment is required, the evaluation of a risk function or a measure of efficiency is necessary. The asymptotic efficiency focuses on the mean squared error (MSE) of prediction which plays the role of the loss function and the average MSE of prediction which plays the role of the expected loss function. Both of these issues should be investigated for the newly developed DIC criterion. Note though that in regard to the asymptotic efficiency we have already obtained in Chapter 3 a lower bound for the mean squared error of prediction. Now we should investigate whether the mean squared error evaluated for the model selected by DIC can attain the known lower bound of prediction. If this can be shown then DIC will be an asymptotically efficient criterion.

Another generalization is the application of the general BHHJ family of measures of divergence to the location model for model selection. The location model (Olkin and Tate, 1961) is an interesting model that describes the joint density of a random vector with both categorical and continuous coordinates. Olkin and Tate consider the problem of the multivariate normal distribution for the continuous component of the density but more general parametric distribution families could be considered. It will be very interesting to define on one hand the general BHHJ measure of divergence in this case and then construct a model selection criterion or generalize the DIC criterion to this special setting. Alternatively an already known divergence could be used for the construction of a proper model selection criterion.

In regard to the goodness of fit tests a number of important issues can be addressed. For example the test statistics proposed can be generalized to cover tests of homogeneity. For example one could test the equality of the measures between functions f_1, g_1 on one hand and f_2, g_2 on the other. In fact such a test could be generalized to r divergences.

Furthermore, effort should be made to improve the asymptotic distribution of the test statistic proposed. As it can be shown from the simulations the asymptotic distribution is not a good approximation of the true distribution of the test statistic. Preliminary simulations (with $n = 500$) show than even if the sample size is large the theoretical powers are still behind the simulated ones.

Let us redefine the general class of BHHJ measures as follows. Let $\mathcal{G}$ be the class of all convex functions Φ on $[0, \infty)$ such that $\Phi(1) = 0$, $\Phi'(1) = 0$ and $\Phi''(1) \neq 0$.

Let f_1 and f_2 be two continuous probability density functions, $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ be two discrete finite probability distributions and μ a given measure. The $(\Phi, a)-$ continuous power divergence family between 2 density functions f_1 and f_2 is defined by:

$$I_{\Phi}^{a}(f_1, f_2) = \int f_2^{1+a} \Phi\left(\frac{f_1}{f_2}\right) d\mu, \;\; a > 0, \Phi \in \mathcal{G},$$

where we assume the conventions $0\Phi\left(0/0\right) = 0$ and $0\Phi\left(u/0\right) = \lim_{u \to \infty} \Phi\left(u\right)/u$, for $u > 0$.

Similarly, the $(\Phi, a)-$discrete power divergence family between two discrete finite probability distributions $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ is defined by

$$d_{\Phi}^{a}(P, Q) = \sum_{j=1}^{m} q_j^{1+a} \Phi\left(\frac{p_j}{q_j}\right), \ a > 0, \Phi \in F.$$

Observe that if

$$\Phi(u) = \Phi_1(u) = u^{1+a} - (1 + \frac{1}{a})u^a + \frac{1}{a}$$

then the BHHJ measure (3.1.1) is obtained, namely,

$$I_{\Phi_1}^{a}(f_1, f_2) = \int f_2^{1+a} \Phi_1\left(\frac{f_1}{f_2}\right) d\mu \equiv I_X^a(f_2, f_1).$$

A similar result is obtained if

$$\Phi(u) = \Phi_2(u) = 1 - (1 + \frac{1}{a})u + \frac{u^{1+a}}{a},$$

namely,

$$I_{\Phi_2}^{a}(f_1, f_2) = \int f_2^{1+a} \Phi_2(\frac{f_1}{f_2}) d\mu \equiv I_X^a(f_1, f_2).$$

The same results hold for the discrete case.

The $(\Phi, a)-$family covers not only the BHHJ measure (Basu et *al.*, 1998) but also the Csiszár's family of measures. Indeed, if we take $\Phi = \phi$ and $a = 0$ then the $(\Phi, a)-$family coincides with Csiszár's measure.

Also, the $(\Phi, a)-$family reduces to the family of Cressie and Read (1984) power divergence family for $a = 0$ and for

$$\Phi(u) = \frac{u^{\lambda+1} - u - \lambda(u-1)}{\lambda(\lambda+1)}, \ \lambda \neq 0, -1.$$

Finally, within the class of the $(\Phi, a)-$family of measures we can introduce and investigate another class of measures which can be considered as a generalization of the family of Csiszár's measures. In particular, consider the case where $\Phi = \phi$ and $a > 0$. Then we define the continuous a-Csizsár family of measures by

$$I_C^a(f_1, f_2) = \int f_2^{1+a} \phi\left(\frac{f_1}{f_2}\right), \ a > 0$$

and the discrete a-Csiszár family of measures by

$$d_c^a(P, Q) = \sum_{j=1}^{m} q_j^{1+a} \phi\left(\frac{p_j}{q_j}\right), \ a > 0.$$

For this last generalized Csiszár's family of measures one could explore its use in developing goodness of fit tests and in particular one could investigate the effect of the index a in improving the power and size of the resulting tests.

Bibliography

1. Aczel, J. (1986). Characterizing information measures: Approaching the end of an era. In *Uncertainty in Knowledge - Based Systems.* Lecture Notes in Computer Science (Eds., B. Bouchon and R. R. Yager), 359–384, Springer - Verlag, New York.

2. Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York, John Wiley.

3. Agresti, A. and Agresti, B. F. (1978). Statistical analysis of qualitative variation. *Sociological Methodology*, **9**, 204-237.

4. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proc. of the* 2^{nd} *Intern. Symposium on Information Theory*, (Petrov B. N. and Csaki F., eds.), 267-281, Akademiai Kaido, Budapest.

5. Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. R. Statist. Soc. B*, **28**, 131-142.

6. Arimoto, S. (1971). Information - theoretical considerations on estimation problems. *Information and Control*, **19**, 181–194.

7. Balakrishnan, V. and Sanghvi, L. D. (1968). Distance between population on the basis of attribute. *Biometrics*, **24**, 859–865.

8. Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.

9. Bhansali, R. J. (1986). Asymptotically efficient selection of the order by the criterion autoregressive transfer function. *Ann. Statist.*, **14**, 315-325.

10. Bhansali, R. J. and Downham D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, **64**, 547-551.

11. Burbea, J. (1984). The Bose-Einstein entropy of degree a and its Jensen difference. *Utilitas Mathematica*, **25**, 225–240.

12. Burbea, J. and Rao, C. R. (1982a). On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, **28**, 489–495.

13. Burbea, J. and Rao, C. R. (1982b). On the convexity of higher order Jensen differences based on entropy functions. *IEEE Transactions on Information Theory*, **28**, 961–963.

14. Burbea, J. and Rao, C. R. (1982c). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. of Multivariate Anal.*, **12**, 575–596.

15. Casella, G. and Berger, R. L. (2001). *Statistical Inference*, 2nd. ed., Duxbury, USA.

16. Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, **3**, 273–304.

17. Cox, D. R. (1970). *The Analysis of Binary Data.* London, Methuen.

18. Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *J. R. Statist. Soc. B*, **5**, 440–464.

19. Cressie, N. and Read, T. R. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer Verlag, New York.

20. Csiszar, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizitat on Markhoffschen Ketten, *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, **8**, 84–108.

21. Csiszar, I. (1977). Information measures: A critical review. *Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, 73–86, Prague, 1974, Academia.

22. Csiszar, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems, *Ann. Statist.*, **19**, 2032-2066.

23. Dawid, A. P. (1998). Coherent measures of discrepancy, uncertainty and dependence, with application to Bayesian predictive experimental design. *Technical Report 139*, Dept. of Statistical Science, University College, London.

24. Dawid, A. P., and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.*, **27**, 65–81.

25. Ferentinos, K. and Papaioannou, T. (1981). New parametric measures of information, *Information and Control*, **51**, 193-208.

26. Ferreri, C. (1980). Hypoentropy and related heterogeneity, divergence and information measures. *Statistica*, **2**, 155–167.

27. Fisher, R. A. (1925). Theory of Statistical Estimation, *Proc. Cambridge Philos. Soc.*, **22**, 700–725.

28. Frieden, B. R. (1988). Applications to optics and wave mechanics of the criterion of maximum Cramer-Rao bound, *J. Modern Optics*, **35**, 1297-1316.

29. Frieden, B. R. (1998). *Physics from Fisher Information - A Unification.* Cambridge University Press, UK.

30. Ghurye, S. G. and Johnson, B. R. (1981). Discrete approximations to the information integral. *The Canadian J. of Statist.*, **9**, 27-37.

31. Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review,* **70**, 419–435.

32. Gini, C. (1912). Variabilita e Mutabilita, Studi Economicoaguaridici della Facotta di Ginrisprudenza dell, Universite di Cagliari III, Parte II.

33. Gokhale, D.V. and Kullback, S. (1978). *The Information in Contingency Tables.* New York, Marcel Dekker.

34. Hunter, D. (2002). *Lecture notes in Asymptotic Tools*, Penn-State University.

35. Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

36. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, Series A, **186**, 453–561.

37. Jones, L. K. and Byrne, C. L. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Trans. Info. Theory*, **36**, 23-30.

38. Kagan, A. M. (1963). On the theory of Fisher's amount of information. *Sov. Math. Dokl.*, **4**, 991-993.

39. Kagan, A. (2001). A discrete version of Stam inequality and a characterization of the Poisson distribution. *J. Statist. Plann. Inference*, **92**, 7-12.

40. Kapur, J. N. (1972). Measures of uncertainty, mathematical programming and physics. *Journal of Indian Society of Agriculture and Statistics*, **24**, 47–66.

41. Karagrigoriou, A. (1997). Asymptotic efficiency of the order selection of a nongaussian AR process. *Statist. Sinica*, **7**, 407-423.

42. Kendall, M. G. (1973). Entropy, probability and information. *International Statistical Review*, **41**, 59–68.

43. Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari*, **4**, 83-91.

44. Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875-890.

45. Kullback, S. (1959). *Information Theory and Statistics.* John Wiley & Sons, New York.

46. Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Math. Statist.*, **22**, 79-86.

47. Lévy, P. (1925). *Calcul des Probabilites.* Gauthiers - Villars, Paris.

48. Liese, F. and Vajda, I. (1987). *Convex Statistical Distances.* Teubner, Leipzig.

49. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, **37**, 145–151.

50. Lindley, D. V. (1956). On a measure of information provided by an experiment. *Annals of Math. Statist.*, **27**, 986-1005.

51. Lindley, D. V. (1961). Dynamic programming and decision theory. *Applied Statistics*, **10**, 39–51.

52. Mathai, A. and Rathie, P. N. (1975), *Basic Concepts in Information Theory.* John Wiley and Sons, New York.

53. Matusita, K. (1951). On the theory of decision functions. *Ann. Inst. Statist. Math.*, **3**, 17-35.

54. Matusita, K. (1964). Distance and decision rules. *Ann. Inst. Statist. Math.*, **16**, 305-320.

55. Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. *Ann. Inst. Statist. Math.*, **19**, 181-192.

56. McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection.* World Scientific Publishing Co.

57. Morales, D., Pardo, L., and Vajda, I. (1997). Some new statistics for testing hypotheses in parametric models. *J. of Multivariate Anal.*, **62 (1)**, 137-168.

58. Nadarajah, S. and Zografos, K. (2003). Formulas for Renyi information and related measures for univariate distributions. *Information Sciences*, **155**, 119-138.

59. Nadarajah, S. and Zografos, K. (2005). Expressions for Renyi and Shannon entropies for bivariate distributions, *Information Sciences*, **170**, 173-189.

60. Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Math. Statist.*, **32**, 448-465.

61. Papaioannou, T. (1985). Measures of information. *Encyclopedia of Statistical Sciences*, Kotz and Johnson, Eds., Wiley, **5**, 391-397.

62. Papaioannou, T. (2001). On distances and measures of information: A case of diversity. *Probability and Statistical Models with Applications*, eds. Charalambides, C. A., Koutras, M. V. and Balakrishnan, N., Chapman & Hall/CRC, 503-514.

63. Papaioannou, T. and Ferentinos, K. (2005). On two forms of Fisher's measure of information. *Commun. in Statist. Theory and Methods*, **34**, 1461-1470.

64. Papaioannou, T. and Kempthorne, O. (1971). On Statistical Information Theory and Related Measures of Information. *Technical Report* No. ARL. 71-0059, Aerospace Research Laboratories, Wright-Patterson A.F.B., Ohio.

65. Pardo, J. A., Pardo, L., Menendez, M. L. (1992). Unified $(r,s)-$entropy as an index of diversity. *J. of the Franklin Instit.*, **329**, 907-921.

66. Pardo, L. (1999). Generalized divergence measures: statistical applications. *Encyclopedia of Microcomputers*, 163–191. Marcel - Dekker, New York.

67. Pardo, L. (2006). *Statistical Inference Based on Divergence Measures.* Chapman & Hall/CRC, Boca Raton.

68. Pardo, L. , Morales, D., Salicru, M. and Menendez, M. L. (1993). R_ϕ^h - divergence statistics in applied categorical data analysis with stratified sampling. *Utilitas Mathematica,* **44**, 145–164.

69. Pardo, M. C. (1999). On Burbea-Rao divergence based goodness of fit tests for multinomial models. *J. of Multivariate Anal.*, **69(1)**, 65-87.

70. Patil, G. P. and Taille, C. (1982). Diversity as a concept and its measurement. *J. of the American Statist. Assoc.*, **77**, 548-567.

71. Pearson, K. (1900). On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine* **50**, 157–172.

72. Rao, C. R. (1958). On an analogue of Cramer-Rao inequality. *Skand. Actuar. Tidskr*, **41**, 5-64.

73. Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

74. Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya A*, **44**, 1-22.

75. Rathie P. N. and Kannappan, P. (1972). A directed-divergence function of type β. *Information and Control*, **20**, 38-45.

76. Renyi, A. (1961). On measures of entropy and information. *Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 547–561.

77. Rukhin, A. L. (1994). Optimal estimator for the mixture parameter by the method of moments and information affinity. *Transactions 12th Prague Conference on Information Theory*, 214–219.

78. Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics.* Dordrecht: D. Reidel.

79. Saks, S. (1937). *Theory of the Integral.* Lwow, Warszawa.

80. Salicru, M. Menendez, M. L. and Pardo, L. (1993). Asymptotic distribution f (h, ϕ)-entropies. *Commun. in Statist. Theory and Methods*, **22**, 7, 2015–2031.

81. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statist.*, **6**, 461-464.

82. Serfling, R. J. (1980). *Approximations Theorems of Mathematical Statistics.* John Wiley, New York.

83. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423.

84. Sharma, B. D. and Mittal, D. P. (1977). New non - additive measure of relative information. *Journal of Combinatory Information and Systems Science*, **2**, 122–133.

85. Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117-126.

86. Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of linear process. *Ann. Statist.*, **8**, 147-164.

87. Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45-54.

88. Shiva, S., Ahmed, N. and Georganas, N. (1973). Order preserving measures of information. *J. Appl. Prob.*, **10**, 666-670.

89. Simpson, E. H. (1949). Measurement of diversity. *Nature* **163**, 688.

90. Soofi, E.S. (1994). Capturing the intangible concept of information. *J. of the American Statist. Assoc.*, **89**, 1243–1254.

91. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B*, **64**, 583-639.

92. Vajda, I. (1973). X^2-divergence and generalized Fisher's information. *Trans.* 6^{th} *Prague Conf. (1971)*, Academia Prague, 873-886.

93. van der Linde, A. (2005). DIC in variable selection. *Statist. Neerlandica*, **59**, 45-56.

94. Wei, C. Z. (1992). On predictive least squares principles. *Ann. of Statist.*, **20**, 1-42.

95. Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, **33**, 1-17.

96. Zografos, K. (1987). Contributions to Statistical Information Theory. *Ph.D. Thesis*, Probability - Statistics and Operational Research Unit, Department of Mathematics, University of Ioannina (in Greek).

97. Zografos, K. (1993). Asymptotic properties of Φ-divergence statistic and applications in contingency tables. *International Journal of Mathematics and Statistical Sciences*, **2**, 5–21.

98. Zografos, K. (1998). On a measure of dependence based on Fisher's information matrix. *Commun. in Statist. Theory and Methods*, **27**, 1715-1728.

99. Zografos, K. (2000). Measures of multivariate dependence based on a distance between Fisher information matrices. *J. Statist. Plann. Inference*, **89**, 91-107.

100. Zografos, K., Ferentinos, K. and Papaioannou, T. (1986). Discrete approximations to the Csiszar, Renyi, and Fisher measures of information. *The Canadian J. of Statist.*, **14(4)**, 355-366.

101. Zografos, K., Ferentinos, K. and Papaioannou, T. (1989). Order preserving property of measures of information. *Commun. in Statist. Theory and Methods*, **18(7)**, 2647-2656.

102. Zografos, K., Ferentinos, K. and Papaioannou, T. (1990). Φ-divergence statistics: Sampling properties, multinomial goodness of fit and divergence tests. *Commun. in Statist. Theory and Methods*, **19(5)**, 1785–1802.