Information theory and identification

Eindhoven
University of Technology
the Netherlands

Department of Electrical Engineering

Information Theory and Identification
by
M.F. Ponomarenko

EUT Report 81-E-122
ISBN 90-6144-122-6
October 1981

Eindhoven University of Technology Research Reports

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering

Eindhoven The Netherlands

INFORMATION THEORY AND IDENTIFICATION

By

M.F. Ponomarenko

EUT Report 81-E-122

ISBN 90-6144-122-6

Eindhoven

October 1981

PREFACE

This report is an attempt to indicate the variety of analytic means suggested by the modern information theory and to give an account of useful and effective applications of information theory in identification. Some new problems in this extensive field, which can be solved within an information-theoretical framework, are also pointed out.

Several information measures are presented in the first part of the report, with special emphasis on their properties and relations to the well-established Shannon entropy. Most of the phenomena dealt with in identification are stochastic. Therefore, restriction has been made to probabilistic information measures involving discrete and continuous probability distributions.

Two kinds of statistical models most frequently used in identification will be distinguished, dependent on whether the probability distributions involved contain unknown parameters or not. Information measures based on such distributions are termed as parametric or non-parametric, respectively. It will be shown that non-parametric information models can be easily extended to parametric distributions, whereas parametric models lose their modelling value when losing the unknown parameters.

Information measures involving discrete probability distributions might seem to be of minor importance for identification, which deals mostly with continuous random variables. The discrete versions, however, are very enlightening for the basic properties of information measures. A transition to continuous analogues is always possible and presents no difficulties, as can be seen from chapters 5 and 6.
The properties of information measures are usually given without proofs. These proofs can be found in the numerous references.

No completeness is claimed; we only discuss those measures which seem to suggest suitable information for identification. On this point it is apparent that new information-theoretic concepts and measures, such as inaccuracy attributed to Kerridge or certainty attributed to Van der Lubbe might lead to new results in estimation and identification. An extension of the former to identification of structure of the model has been demonstrated in chapter 6.

Special attention is paid to the problems where information theory seems to be the best or even the only tool available. One such problem is the choice of the probability distribution related to an estimated process. An extension of the so-called maximum entropy principle, attributed to Jaynes, to parametric statistical models presented in chapter 6, results in a modified decision rule termed minimum information principle, which is based on the Fisher measure of information.

Several formulations of the information-theoretic estimation principle involving Shannon's information measure are also given in the last chapter and numerous contributions to traditional problems of estimation (prediction, filtering, smoothing) and identification made by this principle are discussed. As a conclusion of the present report it appears that the information approach can serve as an appropriate basis for unification, systematization, generalization and further development of the extensive identification field.

This work was done while the author was a visitor in the Measurement and Control Group of the Eindhoven University of Technology. He wishes to express his sincere gratitude to Professor P. Eykhoff for suggesting the problem, holding helpful discussions and giving criticism and full support. The encouragement of Professor P.P. Ornatsky of Kiev Polytechnical Institute, the author's teacher and mentor, is also highly appreciated. Thanks are due to Mr. A.A. van Rede and other members of the Group for their hospitality. The author is indebted to Dr. D.E. Boekee and Dr. A. van den Bos of Delft University of Technology for useful discussions and hospitality. He also thanks librarians Mrs. Henriëtte de Brouwer, Mr. P. van de Ven, Mr. P.S.A. Groot and Ir. I.V. Brůža for their help. For typing as well as for permanent assistance, the author is indebted to Mrs. Barbara Cornelissen.

The cooperation on this project was made possible through the Netherlands Ministry of Education and Sciences, and the Ministry of Higher Special Secondary Education of the USSR.

Dr. M.F. Ponomarenko,
Kiev (Order of Lenin) Polytechnic Institute,
Brest-Litovsky prospekt 39,
KIEV,
USSR

CONTENTS

1. MEASURES OF UNCERTAINTY

Information in an experiment is usually considered as a reduction in uncertainty of the existing knowledge about an event, which is due to observation on this or some related event. Uncertainty is therefore a basic concept of the whole information theory. The first measure of uncertainty termed entropy was introduced by Shannon as early as 1948. The Shannon entropy possesses many useful properties, the most important of which is its (strong) additivity. Moreover, it appears to be the only measure to possess this property, among all possible functions of probabilities satisfying certain intuitively reasonable requirements.

Several generalizations of the Shannon entropy have been developed in the last few decades, which show additivity properties of a different (weaker) kind. One of them is the entropy of order α, introduced by Rényi, which appears to be additive for independent experiments. Another extension due to Havrda and Charvát termed entropy of type β, fails to be additive in the conventional sense, but shows a specific form of additivity, which in many respects makes it even closer to the Shannon entropy as compared with the entropy of order α. A further generalization is given by a so-called entropy of order α and type β, which reduces under certain conditions to the above-mentioned measures. We shall also dwell upon a class of entropies introduced by Arimoto.

Most of the information measures discussed in this project are based on the respective entropy measures. Therefore, we shall pay special attention to the background, definitions and properties of entropies, which will be referred to throughout the report.

1.1 Shannon entropy

Let P denote a set of probabilities $\{p_1,....p_n\}$ with $p_i > 0$, $i = 1,2,...,n$ and $\sum_{i=1}^{n} = 1$, called a complete discrete probability distribution. Each p_i can be considered as a probability of a certain outcome of an experiment, n being the number of all possible outcomes. Uncertainty about the outcome of such an experiment can be expressed by a quantitative measure (Shannon, 1948)

$$H_n(P) = -\sum_{i=1}^{n} p_i \log p_i \qquad (1.1.1)$$

termed Shannon entropy. The Shannon entropy can be regarded as an expectation

$$H_n(P) = E[h(p_i)] = \sum_{i=1}^{n} p_i\, h(p_i), \qquad (1.1.2)$$

where

$$h(p) = -\log p \qquad (1.1.3)$$

is a measure of uncertainty about the outcome of a single event with probability p (or that of a particular outcome of an experiment). The latter can be regarded as a measure of information provided by occurrence of a given event (irrespective of other possible events) and therefore it was originally termed self-information (Wiener, 1948). We prefer the term "self entropy" because h is directly related to the Shannon entropy. The self-entropy given in (1.1.3) is a monotonically decreasing nonlinear function of p. Nonlinearity seems to be well justified by at least two intuitively reasonable requirements. First, the difference in uncertainty about the outcomes of two events with probabilities p and q and a
given difference p-q should be higher for less probable events, i.e. for small p and q. Secondly, uncertainty about outcomes of two independent events is expected to be a sum of uncertainties about the outcome of each event,

$$h(pq) = h(p) + h(q) \text{ for all } p,q \in (0,1] \qquad (1.1.4)$$

Another important property of the self-entropy is non-negativity

$$h(p) \geq 0 \text{ for all } p \in (0,1]. \tag{1.1.5}$$

It can be shown (Luce, 1961; Rényi, 1961; Aczél, 1975) that h defined by (1.1.3) is the only measure possessing the properties given in (1.1.4) and (1.1.5). Adding a normalizing condition, e.g.

$$h(\tfrac{1}{2}) = 1$$

determines the base of logarithm on the right hand side of (1.1.3). The Shannon entropy (1.1.1) can, in turn, be regarded as a self-entropy of one event, whose probability is equal to the mean probability $\overset{\vee}{p}$ of the given distribution P. Setting $p = \overset{\vee}{p}$ in (1.1.3) results in

$$H_n(P) = h(\overset{\vee}{p}) = -\log \overset{\vee}{p}. \tag{1.1.6}$$

Equating the right hand sides of (1.1.1) and (1.1.6) gives and expression for the mean probability in terms of $p_i (i=1,\ldots,n)$

$$\overset{\vee}{p} = \prod_{i=1}^{n} p_i^{\,p_i}. \tag{1.1.7}$$

It follows thus that $\overset{\vee}{p}$ is the weighted geometric mean of p_i with p_i as weights (Aczél, 1975).

A useful measure termed entropy function,

$$f(p) = -p \log p - (1-p)\log(1-p), p \in (0,1] \tag{1.1.8}$$

is a particular case of the Shannon entropy (1.1.1) with $n = 2$ and $p_1 = p$, $p_2 = 1 - p$. It can be considered, hence, as a mean uncertainty about the occurrence or non-occurrence of a single event whose probability is p. The entropy function f(p) can be obtained as a solution of a so-called functional equation of information (Daróczy, 1969) which is a natural consequence of certain reasonable requirements.

Let A and $\overline{A}$ denote the occurrence and non-occurrence of a certain event and $p(A)$, $p(\overline{A})$ denote the corresponding probabilities of occurrence and non-occurrence, so that $p(\overline{A}) = 1 - p(A)$. In order

to derive a reasonable measure of uncertainty concerning A and $\bar{A}$, let us consider another event represented by its occurrence B or non-occurrence $\bar{B}$ with probabilities p(B) and $p(\bar{B})$, respectively. Suppose A and B are independent and mutually exclusive (disjoint) with $p(A) + p(B) \leq 1$. The conditional probability of A given $\bar{B}$ and that of B given $\bar{A}$ are defined by

$$p(A/\bar{B}) = \frac{p(A)}{p(\bar{B})} \tag{1.1.9}$$

and

$$p(B/\bar{A}) = \frac{p(B)}{p(\bar{A})}$$

respectively. Let us require that the measure of uncertainty concerning one event with two possible outcomes A, $\bar{A}$ be a function of the probability p(A) only,

$$H(A) = f(P(A)); \quad H(B) = f(p(B)). \tag{1.1.10}$$

Let

$$H(A/\bar{B}) = p(\bar{B})\, f(p(A/\bar{B})) \quad \text{and} \tag{1.1.11}$$
$$H(B/\bar{A}) = p(\bar{A})\, f(p(B/\bar{A}))$$

be the <u>relative uncertainties</u> of one event with respect to non-occurrence of another one and

$$H(A,B) = H(A) + H(B/\bar{A}) = H(B) + H(A/\bar{B}) \tag{1.1.12}$$

be the <u>joint uncertainty</u> concerning two events.

Substituting (1.1.10) and (1.1.11) in (1.1.12) results in

$$f(p(\bar{A})) + p(\bar{A})\, f(p(B/\bar{A})) = f(p(B)) + p(\bar{B})\, f(p(A/\bar{B})) \tag{1.1.13}$$

Setting $x = p(A)$ and $y = p(B)$, $0 \leq x < 1$, $0 \leq y < 1$, $x + y \leq 1$, in (1.1.13) leads to the functional equation sought (Tverberg, 1958; Aczél, 1975)

$$f(x) + (1-x)\, f\left(\frac{y}{1-x}\right) = f(y) + (1-y)\, f\left(\frac{x}{1-y}\right), \qquad (1.1.14)$$

with a definition domain

$$\{(x,y);\quad 0 \leq x < 1;\quad 0 \leq y < 1;\quad x+y \leq 1\}.$$

The entropy function f as defined by (1.1.8) appears to be a solution of the functional equation (1.1.14) under one additional (boundary) condition given by

$$f(1) = f(0), \qquad (1.1.15)$$

which implies that the entropy of a certain event is equal to the entropy of an impossible event. The Shannon entropy (1.1.1) can be expressed through the entropy function (1.1.8) by

$$H_n(P) = \sum_{i=2}^{n} q_i f\left(\frac{p_i}{q_i}\right), \qquad (1.1.16)$$

where

$$q_i = p_1 + \cdots p_i \qquad (i = 2,\ldots,n).$$

Let us consider two experiments with finite discrete probability distributions of their outcomes $P = \{p_1, \ldots,p_n\}$ and $Q = \{q_1, \ldots,q_m\}$. We can combine them in one single experiment with a probability distribution

$$R=\{r_{11}, r_{12},\ldots,r_{1m}, r_{21}, r_{22},\ldots,r_{2m},\ldots,r_{n1}, r_{n2},\ldots,r_{nm}\},$$

$$\sum_{j=1}^{m} r_{ij} = p_i > 0, \quad \sum_{i=1}^{n} r_{ij} = q_j > 0, \quad \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} = \sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j = 1.$$

The mean uncertainty about the outcome of the first experiment given an outcome of the second experiment is given by the <u>conditional entropy</u>

$$H_{n/m}(P/Q) = -\sum_{j=1}^{m}\sum_{i=1}^{n} r_{ij} \log \frac{r_{ij}}{q_j} \qquad (1.1.17)$$

and the uncertainty concerning the combined experiment is defined by the <u>joint entropy</u>

$$H_{nm}(P,Q) = -\sum_{j=1}^{m}\sum_{i=1}^{n} r_{ij} \log r_{ij} . \qquad (1.1.18)$$

In the following text, we give several properties of the Shannon entropy defined by (1.1.1), (1.1.17) and (1.1.18). The proofs can be found, e.g. in (Aczél, 1975; Mathai, 1975).

1. $H_n(P) \geq 0$
   (non-negativity). (1.1.19)

2. $H_n(p_1,\ldots,p_n) = H_n(p_{k1},\ldots,p_{kn})$,
   where k is any permutation on $\{1,\ldots,n\}$ (symmetry). (1.1.20)

3. $H_{n+1}(p_1,\ldots,p_n,0) = H_n(p_1,\ldots,p_n)$
   (expansibility). (1.1.21)

4. $H_2(\frac{1}{2},\frac{1}{2}) = 1$
   (normality). (1.1.22)

5. $H_n(p_1,\ldots,p_n)$ is a continuous function of p_i,
   $i = 1,\ldots,n$. (1.1.23)

6. $H_n(\frac{1}{n},\ldots,\frac{1}{n}) < H_{n+1}(\frac{1}{n+1},\ldots,\frac{1}{n+1})$ (1.1.24)

   (H_n is a monotonically increasing function of n).

7. $H_n(p_1,\ldots,p_n) \leq H_n(\frac{1}{n},\ldots,\frac{1}{n}) = \log n$ (1.1.25)
   (maximality).

8. In the case n = m (i.e. if the numbers of outcomes in two experiments are equal) it holds

   $$H_n(P) \leq -\sum_{i=1}^{n} p_i \log q_i . \qquad (1.1.26)$$

9. $$H_n(p_1,\ldots,p_n) = H_{n-1}(p_1+p_2,p_3,\ldots,p_n) + (p_1+p_2)H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) \qquad (1.1.27)$$

   (recursivity).

10. $H_{n/m}(P/Q) \leq H_n(P)$, with equality iff $r_{ij} = p_i q_j$, $i = 1,\ldots,n$; $j = 1,\ldots,m$ (i.e. iff the experiments are independent.) (1.1.28)

11. $H_{nm}(P,Q) = H_n(P) + H_{m/n}(Q/P)$ (strong additivity). (1.1.29)

12. $H_{nm}(P,Q) \leq H_n(P) + H_m(Q)$ (subadditivity). (1.1.30)

13. $H_{nm}(P,Q) = H_n(P) + H_m(Q)$, iff $r_{ij} = p_i q_j$ $(i = 1,\ldots,n;\ j = 1,\ldots,m)$ (weak additivity). (1.1.31)

14. $H_2(1,0) = H_2(0,1) = 0$ (decisivity). (1.1.32)

Several characterizations of the Shannon entropy are known. The characterization theorem due to Shannon (1948) states that H_n is the only function satisfying the requirements (1.1.22) (normality), (1.1.23) (continuity), (1.1.24) (monotony) and (1.1.27) (recursivity). It appears, however, that this list of properties was incomplete: an additional requirement of symmetry as given in (1.1.20) is needed. The improved characterizations have been given by Hinčin (1953) and Fadeev (1956). Fadeev's characterization theorem is based on the following postulates: (1.1.20) (symmetry), (1.1.22) (normality), (1.1.23) (continuity) and (1.1.27) (recursivity). The characterization due to Khinchin involves expansibility (1.1.21) normality (1.1.22), continuity as given in (1.1.23) for $n = 2$ only, maximality (1.1.25), strong additivity (1.1.29) and decisivity (1.1.32). The theorem due to Chaundy and McLeod (1960) represents the Shannon entropy as a sum

$$H_n(P) = \sum_{i=1}^{n} g(p_i) \tag{1.1.33}$$

satisfying (1.1.22) (normality) and (1.1.31) (weak additivity), g

being a continuous function of $p \in (0,1]$.

The Shannon measure of uncertainty can be extended to incomplete probability distributions $P = \{p_1,\ldots p_n\}$ with $p_i > 0$ $(i = 1,\ldots,n)$ and $\sum_{i=1}^{n} p_i < 1$ (Rényi, 1961), as defined by

$$H_n(P) = -\frac{1}{\sum_{i=1}^{n} p_i} \sum_{i=1}^{n} p_i \log p_i. \qquad (1.1.34)$$

1.2 Rényi's entropy of order α

A generalization of the Shannon entropy known as the entropy of order α, has been developed by Rényi (1961). It can be obtained by use of the concept of average probability (Daróczy, 1964; Aczél, 1975). Let P be a complete finite discrete probability distribution with

$$p_i > 0 \ (i = 1,\ldots,n) \text{ and } \sum_{i=1}^{n} p_i = 1.$$

The average probability of P can be defined by

$$\breve{p}_n = \phi^{-1}\left[\sum_{i=1}^{n} p_i \ \phi(p_i)\right], \qquad (1.2.1)$$

where $\phi(p)$ is some function of p and ϕ^{-1} is its inverse.

The entropy can be expressed through the average probability as given in (1.1.6). Certainly, different functions ϕ will yield different entropies. Suppose ϕ is strictly monotonic and the related function

$$\phi^*(x) = \begin{cases} x\phi(x); & x \in (0,1] \\ 0; & x = 0 \end{cases} \qquad (1.2.2)$$

is continuous. In that case the average probability is symmetric,

$$\breve{p}_n(p_1,\ldots p_n) = \breve{p}_n(p_{k1},\ldots,p_{kn}), \qquad (1.2.3)$$

expansible,

$$\overset{\vee}{p}_n(p_1,\ldots,p_n) = \overset{\vee}{p}_{n+1}(p_1,\ldots,p_n,0), \tag{1.2.4}$$

and possesses a natural property of a mean value given by

$$\min_{1\leq i\leq n} p_i \leq \overset{\vee}{p}_n(p_1,\ldots,p_n) \leq \max_{1\leq i\leq n} p_i. \tag{1.2.5}$$

In order to generalize the Shannon entropy, we have to weaken the system of desired properties listed in (1.1.19) to (1.1.32). If the generalized measure of uncertainty is expected to be weakly additive as given in (1.1.31), then it can easily be seen from (1.1.6) and (1.1.31) that the following equality should hold

$$\overset{\vee}{p}(P,Q) = \overset{\vee}{p}\,\overset{\vee}{q}, \tag{1.2.6}$$

where P and Q are complete finite discrete probability distributions corresponding to certain independent experiments. The requirements imposed upon ϕ and ϕ^* (strict monotony and continuity, respectively) and (1.2.6) can only be satisfied by two particular functions given in

$$\phi(x) = \log x, \quad x \in (0,1] \tag{1.2.7}$$

and

$$\phi(x) = x^{\alpha-1}, \quad x \in (0,1], \quad \alpha > 0, \ \alpha \neq 1 \tag{1.2.8}$$

(Daróczy, 1964).

Substituting (1.2.7) in (1.2.1) leads, by (1.1.6), to the Shannon entropy as defined in (1.1.1), and with (1.2.8) we obtain in the same manner

$$\overset{\vee}{p}_n = \left(\sum_{i=1}^{n} p_i^{\alpha}\right)^{\frac{1}{\alpha-1}}, \quad 0^{\alpha} = 0, \ \alpha > 0, \ \alpha \neq 1 \tag{1.2.9}$$

and

$${}_{\alpha}H_n(P) = -\log \overset{\vee}{p}_n = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_i^{\alpha}, \quad \alpha > 0, \quad \alpha \neq 1 \tag{1.2.10}$$

called the entropy of order α. It follows thus, that Shannon entropy and Rényi's entropy of order α are the only measures of uncertainty satisfying (1.1.31) (weak additivity condition). Note that

the Rényi's entropy of order α reduces to the Shannon entropy in one particular (limiting) case,

$$\lim_{\alpha \to 1} {}_{\alpha}H_n(P) = H_n(P) \tag{1.2.11}$$

For $\alpha \to 0$, we have

$${}_{\alpha}H_n(P) = \log n, \tag{1.2.12}$$

which implies that the entropy of order α is equivalent to the Hartley's measure of uncertainty, depending only on the number of events, when α vanishes. An extension of (1.2.10) to incomplete probability distributions is given by

$${}_{\alpha}H_n(P) = \frac{1}{1-\alpha} \log \frac{\sum_{i=1}^{n} p_i^{\alpha}}{\sum_{i=1}^{n} p_i}, \tag{1.2.13}$$

with $P = \{p_1,\ldots,p_n\}$, $p_i > 0$ $(i=1,\ldots,n)$ and

$\sum_{i=1}^{n} p_i < 1$ (Rényi, 1961).

Let $P = \{p_1,\ldots,p_n\}$ and $Q = \{q_1,\ldots,q_n\}$ denote two complete discrete probability distributions. Suppose

$p_{ij} \in P_j$ $(i=1,\ldots,n;\ j=1,\ldots,m)$ is a conditional probability of an outcome of the first experiment, corresponding to P, with respect to a certain outcome of the second experiment, corresponding to Q,

$q_{ji} \in Q_i$ $(i = 1,\ldots,n;\ j = 1,\ldots,m)$ is a conditional probability of an outcome of the second experiment with respect to a certain outcome of the first on and $r_{ij} = q_j p_{ij} = p_i q_{ji}$ is a joint probability of an outcome of a compound experiment consisting of the performance of both the first and the second one.

The conditional and joint entropies of order α can be obtained by use of the corresponding average probabilities. Replacing p_i in (1.2.9) by $p_{ij}(i=1,\ldots,n)$ results in

$$ {}_{\alpha}\overset{\vee}{p}_n(j) = \Big(\sum_{i=1}^{n} p_{ij}^{\alpha} \Big)^{\frac{1}{\alpha-1}} , \quad j = 1,\ldots,m, \tag{1.2.14} $$

which can be considered as the average probability of the conditional distribution P_j, i.e. the average probability of the distribution P given a certain outcome of the second experiment corresponding to Q. Taking expectation of ${}_{\alpha}\overset{\vee}{p}_{n(j)}$ with respect to the distribution Q leads to the average conditional probability of P given Q,

$$ {}_{\alpha}\overset{\vee}{p}_{n/m} = E\big[{}_{\alpha}\overset{\vee}{p}_{n(j)}\big] = \sum_{j=1}^{m} q_j \Big(\sum_{i=1}^{n} p_{ij}^{\alpha} \Big)^{\frac{1}{\alpha-1}}. \tag{1.2.15} $$

In the same manner we obtain the average probability of Q given P,

$$ {}_{\alpha}\overset{\vee}{p}_{m/n} = E\big[{}_{\alpha}\overset{\vee}{p}_{m(i)}\big] = \sum_{i=1}^{n} p_i \Big(\sum_{j=1}^{m} q_{ji}^{\alpha} \Big)^{\frac{1}{\alpha-1}} . \tag{1.2.16} $$

Taking the logarithm of the right hand side in (1.2.15), we obtain

$$ {}_{\alpha}H_{n/m}(P/Q) = -\log \sum_{i=1}^{n} q_j \Big(\sum_{j=1}^{m} p_{ij} \Big)^{\frac{1}{\alpha-1}} \tag{1.2.17} $$

which is the conditional entropy of order α of the distribution P with respect to the distribution Q. In the same manner we can derive the conditional entropy of order α for Q with respect to P,

$$ {}_{\alpha}H_{m/n}(Q/P) = -\log \sum_{i=1}^{n} p_i \Big(\sum_{j=1}^{m} q_{ji}^{\alpha} \Big)^{\frac{1}{\alpha-1}} \tag{1.2.18} $$

A different definition of the conditional entropy of order α arises if we interchange in (1.2.15) the operation of expectation

and that of raising to the power $\frac{1}{\alpha-1}$, namely

$$ {}_{\alpha}p'_{n/m} = \left[\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{ij}^{\alpha} \right]^{\frac{1}{\alpha-1}} \qquad (1.2.19) $$

(Van der Lubbe, 1981). The corresponding conditional entropy attains the form

$$ {}_{\alpha}H'_{n/m}(P/Q) = \frac{1}{\alpha-1} \log \left[\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{ij}^{\alpha} \right]. \qquad (1.2.20) $$

Expressions for ${}_{\alpha}\check{p}'_{m/n}$ and ${}_{\alpha}H'_{m/n}(Q/P)$ can be found analogously. In the limiting case, $\alpha \to 1$, both ${}_{\alpha}H_{n/m}(P/Q)$ and ${}_{\alpha}H'_{n/m}(P/Q)$ reduce to the Shannon conditional entropy given in (1.1.17). Other definitions of the conditional entropy of order α have been introduced in (Aczél, 1963) and (Arimoto, 1977).

For a compound experiment consisting of the performance of two experiments, corresponding to P and Q, the average probability of its outcome is given by

$$ {}_{\alpha}\check{r}_{nm} = \left[\sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij}^{\alpha} \right]^{\frac{1}{\alpha-1}} \qquad (1.2.21) $$

which is an analogue of (1.2.9). Taking logarithm of the right hand side in (1.2.21), we obtain the following expression for the joint entropy of order α

$$ {}_{\alpha}H_{nm}(P,Q) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij}^{\alpha}. \qquad (1.2.22) $$

For independent P and Q we have $r_{ij} = p_i q_j$ $(i=1,\ldots,n;$ $j=1,\ldots,m)$, and (1.2.22) reduces to

$$
\begin{aligned}
{}_{\alpha}H_{nm}(P,Q) &= \frac{1}{1-\alpha}\log\sum_{i=1}^{n} p_i^{\alpha} + \frac{1}{1-\alpha}\log\sum_{j=1}^{m} q_j^{\alpha} \\
&= {}_{\alpha}H_n(P) + {}_{\alpha}H_m(Q), \qquad (1.2.23)
\end{aligned}
$$

which shows the weak additivity of the entropy of order α.

It can easily be seen that the properties (1.1.19) to (1.1.25) and (1.1.32) of the Shannon entropy also hold for the entropy of order α. The property defined in (1.1.30) (subadditivity) holds for $\alpha=1$ only. Some other properties, which are typical for the entropy of order α, can be expressed as follows (Van der Lubbe, 1981).

1. $$\lim_{\alpha\to\infty} {}_{\alpha}H_n(P) = -\log \max_{1\le i\le n} (p_1,\ldots,p_i,\ldots,p_n). \qquad (1.2.24)$$

2. $$\alpha_2 > \alpha_1 \quad \text{implies} \quad {}_{\alpha_2}H_n(P) \le {}_{\alpha_1}H_n(P) \qquad (1.2.25)$$

   (i.e. ${}_{\alpha}H_n(P)$ is a decreasing function of α), with equality for $P = \{\frac{1}{n},..,\frac{1}{n}\}$ and for $P = \{0,\ldots,0,1,0,\ldots,0\}$

3. (Concavity properties) (1.2.26)
   a) For $\alpha \in (0,1]$, ${}_{\alpha}H_n(P)$ is strictly concave with respect to P.
   b) For $\alpha \in (0,2]$ and $n = 2$, ${}_{\alpha}H_n(P)$ is also concave with respect to P.
   c) For $\alpha > 2$ and $n > 2$, ${}_{\alpha}H_n(P)$ is neither concave nor convex with respect to P.
   d) For every $\alpha > 1$, there exists an n' such that ${}_{\alpha}H_n(P)$ is neither concave nor convex with respect to P for all $n > n'$.

The concavity properties have been proved by Ben-Bassat and Raviv (1978).

1.3 Entropy of type β

Another generalisation of Shannon's entropy termed entropy of type β is due to Havrda and Charvát (1967). It can be obtained by a certain generalization of the functional equation (1.1.14). Replacing the definition (1.1.11) of the relative uncertainty by

$$H(A/\bar{B}) = p^{\beta}(\bar{B})\, f\,(P(A/\bar{B}));\quad H(B/\bar{A}) = p^{\beta}(\bar{A})\, f\,(p(B/\bar{A})) \qquad (1.3.1)$$

results in a different functional equation,

$$f(x) + (1-x)^{\beta} f\left(\frac{y}{1-x}\right) = f(y) + (1-y)^{\beta} f\left(\frac{x}{1-y}\right). \qquad (1.3.2)$$

The real valued function f, defined in $[0,1]$ and satisfying (1.3.2) under the boundary conditions

$$f(0) = f(1) \qquad (1.3.3)$$

and

$$f(\tfrac{1}{2}) = 1 \qquad (1.3.4.)$$

is called entropy function of type β. This function is given by

$$f_{\beta}(x) = \frac{1}{2^{1-\beta}}\,(x^{\beta} + (1-x)^{\beta} - 1),\quad \beta \neq 1 \qquad (1.3.5)$$

(Daróczy, 1970). Analogously to (1.1.16) we obtain an expression for the entropy of type β:

$$H_n^{\beta}(P) = \sum_{i=1}^{n} q_i^{\beta} f_{\beta}\left(\frac{p_i}{q_i}\right),\quad q_i = p_1 + \ldots + p_i,\quad (i=2,\ldots,n). \qquad (1.3.6)$$

Substituting (1.3.5) in (1.3.6) results in an explicit definition

$$H_n^{\beta}(P) = \frac{1}{1-2^{1-\beta}}\left(1 - \sum_{i=1}^{n} p_i^{\beta}\right),\quad \beta > 0,\quad \beta \neq 1, \qquad (1.3.7)$$

with convention $0^{\beta} = 0$ $(\beta \neq 0)$.

The entropies of order α and of type β are related to each other by

$$H_n^{\beta}(P) = \frac{1}{2^{1-\beta}-1}\left(2^{(1-\beta)\,{}_{\beta}H_n(P)}-1\right), \qquad (1.2.8)$$

and

$${}_{\alpha}H_n(P) = \frac{1}{1-\alpha}\log\left((2^{1-\alpha}-1)\,H_n^{\alpha}(P)+1\right) \qquad (1.3.9)$$

which follow from (1.2.10) and (1.3.7).

It can be seen from the list of the properties of $H_n{}^{\beta}(P)$ given below, that the entropy of type β shows even more resemblance to the Shannon entropy than the entropy of order α.

1. Non-negativity, as defined by (1.1.19)
2. Symmetry, as defined by (1.1.20)
3. Expansibility, as defined by (1.1.21)
4. Normality, as defined by (1.1.22)
5. Continuity, as defined by (1.1.23)
6. Monotony, as defined by (1.1.24)
7. Maximality, as defined by (1.1.25), with $\frac{1-n^{1-\beta}}{1-2^{1-\beta}}$ in place of $\log n$.
8. Recursivity of type β:

$$H_n^{\beta}(p_1,\cdots,p_n) = H_{n-1}^{\beta}(p_1+p_2,p_3,\cdots,p_n) + (p_1+p_2)^{\beta}\,H_2^{\beta}\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right) \qquad (1.3.10)$$

with $p_i \geq 0$ $(i=1,\cdots,n)$ and $p_1 + p_2 \neq 0$.

9. Strong additivity of type β

$$H_{nm}^{\beta}(p_1q_{11},\cdots,p_1q_{m1},p_2q_{12},\cdots,p_2q_{m2},\cdots,p_nq_{1n},\cdots,p_nq_{mn})$$

$$= H_n^{\beta}(p_1,\ldots,p_n) + \sum_{i=1}^{n} p_i^{\beta} H_m^{\beta}(q_{1i},q_{2i},\ldots q_{mi}), \qquad (1.3.11)$$

with

$$p_i > 0 \ (i=1,\ldots,n), \ \sum_{i=1}^{n} p_i \leq 1, \ q_j > 0 \ (j=1,\ldots,m) \text{ and } \sum_{j=1}^{m} q_j \leq 1.$$

10. Weak additivity of type β (non-additivity)

$$H_{nm}^{\beta}(p_1q_1,\ldots,p_1q_m,\ldots p_nq_1,\ldots,p_nq_m) \qquad (1.3.12)$$

$$= H_n^{\beta}(p_1,\ldots,p_n) + H_m^{\beta}(q_1,\ldots,q_m) + (2^{1-\beta}-1)H_n^{\beta}(p_1,\ldots,p_n)H_m^{\beta}(q_1,..,q_m)$$

11. Decisivity, as defined by (1.1.32).

12. Concavity with respect to P for $\beta > 1$ and convexity for $\beta < 1$ (Sharma, 1973). (1.3.13)

13. Limiting properties (1.3.14)

a) $\lim_{\beta\to 0+} H_n^{\beta}(P) = n - 1.$

b) $\lim_{\beta\to 1} H_n^{\beta}(P) = H_n(P).$

c) $\lim_{n\to\infty} H_n^{\beta}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \begin{cases} \frac{1}{1-2^{1-\beta}}; & \beta > 1, \\ \infty & \beta \in (0,1). \end{cases}$

Taking into consideration (1.3.11), a reasonable definition of the <u>conditional entropy of type β</u> seems to be

$$H_{m/n}^{\beta}(Q/P) = \sum_{i=1}^{n} p_i^{\beta} H_m^{\beta}(q_{1i}, q_{2i},\ldots,q_{mi}) \qquad (1.3.15)$$

and

$$H_{n/m}(P/Q) = \sum_{j=1}^{m} q_j^{\beta} H_n^{\beta}(p_{1j},p_{2j},\ldots,p_{nj}). \qquad (1.3.16)$$

In a limiting case, when $\beta \to 1$, the conditional entropy of type β reduces to the Shannon conditional entropy defined by (1.1.17). The following relation

$$H^{\beta}_{n/m}(P/Q) \leq H^{\beta}_{n}(P) \tag{1.3.17}$$

implies that the conditional entropy of type β cannot exceed the marginal entropy of type β.

The joint entropy of type β is defined by

$$H^{\beta}_{nm}(P,Q) = \frac{1}{1-2^{1-\beta}}\left(1-\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}^{\beta}\right), \tag{1.3.18}$$

which follows from (1.3.7) after replacing p_i by r_{ij}. In the limiting case, when $\beta \to 1$, the joint entropy of type β reduces to the Shannon joint entropy given in (1.1.18). The property (1.3.11) implying strong additivity of type β can be expressed by (1.3.15) and (1.3.18) as

$$H^{\beta}_{nm}(P,Q) = H^{\beta}_{n}(P) + H^{\beta}_{m/n}(Q/P). \tag{1.3.19}$$

For independent distributions P,Q, (1.3.19) reduces to

$$H^{\beta}_{nm}(P,Q) = H^{\beta}_{n}(P) + H^{\beta}_{m}(Q) + (2^{1-\beta}-1)\, H^{\beta}_{n}(P)\, H^{\beta}_{m}(Q), \tag{1.3.20}$$

which is another expression for weak additivity given in (1.3.12). As $\beta \to 1$, this relation attains the form of (1.1.31). On account of (1.3.17) and (1.3.19) we also have

$$H^{\beta}_{nm}(P,Q) \leq H^{\beta}_{n}(P) + H^{\beta}_{m}(Q), \tag{1.3.21}$$

which shows that the entropy of type β is subadditive (cf. (1.1.30)). The entropy of type β can be expressed as an ordinary sum

$$H^{\beta}_{n}(P) = \sum_{i=1}^{n} f(p_i), \tag{1.3.22}$$

with

$$f(p_i) = \frac{p_i^{\beta} - p_i}{2^{1-\beta} - 1}, \quad \beta > 0, \ \beta \neq 1. \tag{1.3.23}$$

The characterization theorem due to Daróczy (1970) is based on three postulates defined by (1.1.20) (symmetry), (1.1.22) (normality) and (1.3.10) (recursivity of type β). Another characterization can be found in (Forte, 1973; Mathai, 1975).

1.4 Entropy of order α and type β

The entropy of type β given in (1.3.7) can be represented by a weighted sum

$$H_n^{\beta}(P) = \sum_{i=1}^{n} p_i h^{\beta}(p_i), \tag{1.4.1}$$

with

$$h^{\beta}(p) = \frac{1-p^{\beta-1}}{1-2^{1-\beta}}, \quad \beta \neq 1, \quad \beta > 0. \tag{1.4.2}$$

Sharma and Mittal (1975) have shown that the function $h^{\beta}(p)$ as defined in (1.4.2), called also self-entropy of type β, is the only function satisfying the following postulates

1. $h^{\beta}(p)$ is continuous in $(0,1]$, (1.4.3)
2. $h^{\beta}(pq) = h^{\beta}(q) + h^{\beta}(q) + \lambda\, h^{\beta}(p)\, h^{\beta}(q), \quad \lambda \neq 0,$ (1.4.4)
3. $h^{\beta}(\tfrac{1}{2}) = 1,$ (1.4.5)

which correspond to the properties of the Wiener's self-entropy given in (1.1.3). The generalization consists in (1.4.4) implying strong additivity of type β. Following Sharma and Mittal, let us consider the entropy as a generalized average self-entropy of type β,

$$\phi_{H_n}(P) = \phi^{-1}\left[\sum_{i=1}^{n} p_i \phi\,(h^{\beta}(p_i))\right], \tag{1.4.6}$$

where φ is such a strictly monotonic continuous function that $\phi_{H_n}(P)$ satisfies the requirement of weak additivity as defined by (1.3.20). These conditions admit two solutions given by

$${}_1H_n^{\beta}(P) = \frac{1-2^{(\beta-1)\sum_i^n p_i \log p_i}}{1-2^{1-\beta}}, \quad \beta \neq 1, \quad \beta > 0 \tag{1.4.7}$$

and

$$ {}_{\alpha}H_n^{\beta}(P) = \frac{1-\left[\sum_{i=1}^{n} p_i^{\alpha}\right]^{\frac{\beta-1}{\alpha-1}}}{1-2^{1-\beta}}, \quad \beta \neq 1,\ \beta > 0;\ \alpha \neq 1,\ \alpha > 0, \tag{1.4.8} $$

called the entropy of order 1 and type β and the entropy of order α and type β, respectively.

Being a further generalization, this measure possesses fewer properties of the Shannon entropy than the entropy of order α or the entropy of type β. Taking into consideration that ${}_1H_n^{\beta}(P)$ can be regarded as a particular kind of ${}_{\alpha}H_n^{\beta}(P)$ when $\alpha \to 1$, most of their properties can be expressed in terms of ${}_{\alpha}H_n^{\beta}(P)$ only. In what follows we give a list of the basic properties of ${}_{\alpha}H_n^{\beta}(P)$.

1. Non-negativity, as given in (1.1.19).
2. Symmetry, as given in (1.1.20).
3. Expansibility, as given in (1.1.21)
4. Normality, as given in (1.1.22)
5. Continuity (in P), as given in (1.1.23).
6. Monotony (in n), as given in (1.1.24).
7. Maximality

$$ {}_{\alpha}H_n^{\beta}(P) \leq {}_{\alpha}H_n^{\beta}\left(\frac{1}{n},\ldots,\frac{1}{n}\right) = \frac{1-n^{1-\beta}}{1-2^{1-\beta}}, \quad \alpha > 1. \tag{1.4.9} $$

8. Monotony (in α)
   For a fixed β, ${}_{\alpha}H_n^{\beta}(P)$ is a monotonically decreasing function of α and thus

$$ {}_{\alpha}H_n^{\beta}(P) \leq {}_1H_n^{\beta}(P), \quad \alpha > 1. \tag{1.4.10} $$

9. Weak additivity of type β

$$ {}_{\alpha}H_{nm}^{\beta}(p_1q_1,\ldots,p_1q_m,p_2q_1,\ldots p_2q_m,\ldots,p_nq_1,\ldots,p_nq_m) $$
$$ = {}_{\alpha}H_n^{\beta}(p_1,\ldots,p_n) + {}_{\alpha}H_m^{\beta}(q_1,\ldots,q_m) + $$

$$+ (2^{1-\beta}-1)\ {}_{\alpha}H_n^{\beta}(p_1,\ldots,p_n) \qquad (1.4.11)$$

10. Decisivity, as given (1.1.32).

11. Limiting properties

a) $$\lim_{n\to\infty}\ {}_{\alpha}H_n^{\beta}\left(\frac{1}{n},\ldots,\frac{1}{n},\right) = \frac{1}{2^{1-\beta}-1}, \quad \beta \neq 1. \qquad (1.4.12)$$

b) $$\lim_{\alpha\to 1}\ {}_{\alpha}H_n^{\beta}(P) = {}_{1}H_n^{\beta}(P) \qquad (1.4.13)$$

c) $$\lim_{\beta\to 1}\ {}_{\alpha}H_n^{\beta}(P) = {}_{\alpha}H_n(P), \qquad (1.4.14)$$

where ${}_{\alpha}H_n(P)$ is the entropy of order α, as defined by (1.2.10).

d) $$\lim_{\alpha\to 1}\left[\lim_{\beta\to 1}\ {}_{\alpha}H_n^{\beta}(P)\right] = \lim_{\beta\to 1}\ {}_{1}H_n^{\beta}(P) = H_n(P), \qquad (1.4.15)$$

where $H_n(P)$ is the Shannon entropy as defined by (1.1.1).

Setting $\alpha = \beta \neq 1$ in (1.4.8) gives the entropy of type β, ${}^{\beta}H_n(P)$, defined by (1.3.7). In general, the following relations hold

$${}_{1}H_n^{\beta}(P) = \frac{1-2^{(1-\beta)H_n(P)}}{1-2^{1-\beta}}, \quad \beta \neq 1, \quad \beta \neq 0, \qquad (1.4.16)$$

$${}_{\alpha}H_n^{\beta}(P) = \frac{1-2^{(1-\beta)\,{}_{\alpha}H_n(P)}}{1-2^{1-\beta}}, \quad \beta \neq 1, \quad \alpha \neq 1, \ \alpha \neq 0, \ \beta \neq 0, \qquad (1.4.17)$$

$$ {}_{\alpha}H^{\beta}_{n}(P) = \frac{1-\left[(2^{1-\alpha}-1)\, H^{\alpha}_{n}(P) + 1\right]^{\frac{\beta-1}{\alpha-1}}}{1 - 2^{1-\beta}}, \qquad (1.4.18) $$

which follow directly from the definitions (1.1.1), (1.2.10), (1.3.7), (1.4.7) and (1.4.8).

Analogously to (1.4.6), the conditional entropy of order α and type β can be defined as an average

$$ {}^{\emptyset}H_{n/m}(P/Q) = \emptyset^{-1}\Big\{ \sum_{j=1}^{m} q_j \, \emptyset \left[{}^{\emptyset}H_{n}(P/j) \right] \Big\}, \qquad (1.4.19) $$

with

$$ {}^{\emptyset}H_{n}(P/j) = \emptyset^{-1}\Big\{ \sum_{i=1}^{n} p_{ij} \, \emptyset \left[h^{\beta}(p_{ij}) \right] \Big\}, \qquad (1.4.20) $$

$\emptyset$ being the same monotonic and continuous function as in (1.4.6).

Again, two possible functions $\emptyset$ admit two different measures given by

$$ {}_{1}H^{\beta}_{n/m}(P/Q) = \frac{1-2^{(1-\beta)\, H_{n/m}(P/Q)}}{1-2^{1-\beta}}, \quad \beta \neq 1, \quad \beta > 0 \qquad (1.4.21) $$

where $H_{n/m}(P/Q)$ is the Shannon conditional entropy as defined in (1.1.17) and

$$ {}_{\alpha}H^{\beta}_{n/m}(P/Q) = \frac{1 - \Big[\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{ij}^{\alpha} \Big]^{\frac{\beta-1}{\alpha-1}}}{1-2^{1-\beta}}, \quad \begin{array}{l} \beta \neq 1, \ \beta > 0; \\ \alpha \neq 1, \ \alpha > 0. \end{array} \qquad (1.4.22) $$

Expressions for ${}_{1}H^{\beta}_{m/n}(Q/P)$ and ${}_{\alpha}H^{\beta}_{m/n}(Q/P)$ can be obtained from (1.4.21) and (1.4.22) after replacing p_i by q_j and p_{ij} by q_{ji}.

The joint entropy of order α and type β is given by

$$ {}_1H^{\beta}_{nm}(P,Q) = \frac{1-2^{(\beta-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij} \log r_{ij}}{1 - 2^{1-\beta}}, \quad \beta \neq 1, \ \beta > 0 \tag{1.4.23} $$

and

$$ {}_{\alpha}H^{\beta}_{nm}(P,Q) = \frac{1 - \left(\sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij}^{\alpha} \right)^{\frac{\beta-1}{\alpha-1}}}{1 - 2^{1-\beta}}, \quad \beta \neq 1, \ \beta > 0, \alpha \neq 1, \ \alpha > 0, \tag{1.4.24} $$

which follow from (1.4.7) and (1.4.8) by substituting r_{ij} for p_i. The marginal, conditional and joint entropies of order α and type β are related by

$$ {}_1H^{\beta}_{nm}(P,Q) = {}_1H^{\beta}_{n}(P) + {}_1H^{\beta}_{m/n}(Q/P) - (1-2^{1-\beta}) \, {}_1H^{\beta}_{n}(P) \, {}_1H^{\beta}_{m/n}(Q/P), \tag{1.4.25} $$

$$ {}_{\alpha}H^{\beta}_{n}(P) + {}_{\alpha}H^{\beta}_{m/n}(Q/P) - (1-2^{1-\beta}) \, {}_{\alpha}H^{\beta}_{n}(P) \, {}_{\alpha}H^{\beta}_{m/n}(Q/P) = $$

$$ = \frac{1-2^{\,{}_{\alpha}H_n(P) + {}_{\alpha}H_{m/n}(Q/P)}}{1-2^{1-\beta}} \tag{1.4.26} $$

where ${}_{\alpha}H_n(P)$ and ${}_{\alpha}H_{m/n}(Q,P)$ are the marginal and conditional entropies of order α and

$$ {}_1H^{\beta}_{n}(P) \geq {}_1H^{\beta}_{n/m}(P/Q), \tag{1.4.27} $$

$$ {}_{\alpha}H^{\beta}_{n}(P) \geq {}_{\alpha}H^{\beta}_{n/m}(P/Q), \tag{1.4.28} $$

$$ {}_1H_{nm}^{\beta}(p,Q) \leq {}_1H_n^{\beta}(P) + {}_1H_m^{\beta}(Q) - (1-2^{1-\beta})\, {}_1H_n^{\beta}(P)\, {}_1H_m^{\beta}(Q) $$

$$ b \neq 1,\ \beta > 0, \qquad (1.4.29) $$

with equality in (1.4.27) to (1.4.29) for independent P and Q. The proofs of these relations are given in (Sharma, 1975). Note that (1.4.25) implies the strong additivity property of type β for ${}_1H_n^{\beta}(P)$.

1.5 Arimoto's entropies

The entropies discussed so far can be expressed either as a (weighted) sum or a certain average of self-entropies or as a weighted sum of certain entropy functions. A different approach to design of the measures of uncertainty has been developed by Arimoto (1971). Arimoto's entropies are given by a class of functions

$$ {}_fH_n(P) = \inf_{\tilde{P}} \sum_{i=1}^{n} p_i f(\tilde{p}_i), \qquad (1.5.1) $$

where f is a non-negative real valued function having a continuous derivative and defined in $(0,1]$; $P = \{p_1,\ldots p_n\}$, $\sum_{i=1}^{n} p_i = 1$, $p_i > 0$ $(i = 1,\ldots,n)$ and $\tilde{P} = \{\tilde{p}_1,\ldots,\tilde{p}_n\}$, $\sum_{i=1}^{n} \tilde{p}_i = 1$, $\tilde{p}_i > 0$ $(i = 1,\ldots,n)$.

A particular entropy measure can be derived from (1.5.1) by performing operation inf for a given function f. All of the entropies belonging to the class given in (1.5.1) possess the following properties (Arimoto, 1971).

1. Non-negativity, as defined by (1.1.19).
2. Symmetry, as defined by (1.1.20).
3. Expansibility, as defined by (1.1.21).
4. Continuity in P, as defined by (1.1.23).

5. Maximality

$$ {}_{f}H_{n}(P) \leq {}_{f}H_{n}\left(\frac{1}{n},\ldots,\frac{1}{n}\right) = f\left(\frac{1}{n}\right), \qquad (1.5.2) $$

provided f(p) is a convex function of p.

6. Concavity with respect to P. (1.5.3)

7. Inequality

$$ {}_{f}H_{n}(P) \leq \sum_{i=1}^{n} p_i f(p_i) \qquad (1.5.4) $$

(cf. (1.1.26)).

For n = 2, the equality in (1.5.4) holds for an infinite family of functions f, including f(p) = Clog p, where C is a non-positive constant. For n > 3, the latter is the only function yielding the said equality. In a particular case, when the function f attains the form

$$ f(p) = \frac{R}{R-1}\left(1-p^{\frac{R-1}{R}}\right), \quad R \neq 1,\ R > 0, \qquad (1.5.5) $$

the expression (1.5.1) reduces to

$$ {}_{R}H_{n}(P) = \frac{R}{R-1}\left[1-\left(\sum_{i=1}^{n} p_i^{R}\right)^{\frac{1}{R}}\right], \quad R \neq 1, \quad R > 0. \qquad (1.5.6) $$

This entropy was suggested by Arimoto (1971) and extensively studied by Boekee and Van der Lubbe (1980). In addition to the properties 1 to 7 above, the so-called R-norm entropy measure (1.5.6) possesses

1. Minimality

$$ {}_{R}H_{n}(P) \geq {}_{R}H_{n}(0,\ldots,0,1,0,\ldots,0) = 0. \qquad (1.5.7) $$

2. Monotony, as defined by (1.1.24).

3. Pseudo-additivity

$$ {}_{R}H_{nm}(P,Q) = {}_{R}H_{n}(P) + {}_{R}H_{m}(Q) - \frac{R-1}{R}\,{}_{R}H_{n}(P)\;{}_{R}H_{m}(Q). \qquad (1.5.8) $$

4. Decisivity, as given in (1.1.32).

5. Continuity with respect to the real constant R. (1.5.9)

6. Limiting properties

a) $$\lim_{R\to\infty} {}_RH_n(P) = 1 - \max_i p_i, \quad i = 1,\ldots,n. \tag{1.5.10}$$

b) $$\lim_{R\to 1} {}_RH_n(P) = -\sum_{i=1}^{n} p_i \log p_i = H_n(P), \tag{1.5.11}$$

where $H_n(P)$ is the Shannon entropy given in (1.1.1).

The R-norm entropy is related to the entropy of order α by

$$ {}_RH_n(P) \leq {}_\alpha H_n(P), \quad R > 1, \ \alpha = R; \tag{1.5.12}$$

$$ {}_RH_n(P) \geq {}_\alpha H_n(P), \quad R \in (0,1), \ \alpha = R \tag{1.5.13}$$

and

$$ {}_RH_n(P) = \frac{R}{R-1}\left(1-2^{\frac{1-R}{R}\,{}_\alpha H_n(P)}\right), \quad \alpha = R. \tag{1.5.14}$$

The relation to the entropy of type β is given by

$$ {}_RH_n(P) = \frac{R}{R-1}\left\{1-\left[-(1-2^{1-R})\,H_n^{\beta}(P)\right]^{\frac{1}{R}}\right\}, \quad \beta = R. \tag{1.5.15}$$

The <u>conditional R-norm entropy</u> can be defined as an expectation of

$$ {}_RH_n(P/j) = \frac{R}{R-1}\left[1-\left(\sum_{i=1}^{n} p_{ij}^{R}\right)^{\frac{1}{R}}\right], \quad R \neq 1, \ R > 0,$$

$j = 1,\ldots,m$, which follows from (1.5.6) by substituting p_{ij} for p_i. This approach results in

$$ {}_{R}H_{n/m}(P/Q) = E_{Q}\left[{}_{R}H_{n}(P/j)\right] $$

$$ = \frac{R}{R-1}\left[1-\sum_{j=1}^{m} q_{j}\left(\sum_{i=1}^{n} p_{ij}^{R}\right)^{\frac{1}{R}}\right],\ R \neq 1,\ R > 0. \qquad (1.5.16) $$

Another conditional R-norm entropy can be obtained by interchanging in (1.5.15) the operation of mathematical expectation with respect to Q and the operation of raising the power 1/R (Boekee, 1980),

$$ {}_{R}H'_{n/m}(P/Q) = \frac{R}{R-1}\left[1-\left(\sum_{j=1}^{m} q_{j}\sum_{i=1}^{n} p_{ij}^{R}\right)^{\frac{1}{R}}\right],\ R \neq 1.\ R > 0. \qquad (1.5.17) $$

Both (1.5.15) and (1.5.16) satisfy the following desired requirement

$$ {}_{R}H_{n/m}(P/Q) \leq {}_{R}H_{n}(P), \qquad (1.5.18) $$

with equality iff P and Q are independent.

The joint R-norm entropy

$$ {}_{R}H_{nm}(P,Q) = \frac{R}{R-1}\left[1-\left(\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}^{R}\right)^{\frac{1}{R}}\right],\quad R \neq 1,\quad R > 0. \qquad (1.5.19) $$

can be derived from (1.5.6) by substituting r_{ij} for p_i.

2. MEASURES OF DIVERGENCE

In this chapter we shall discuss measures of dissimilarity between discrete probability distributions called divergence measures, which have been developed within an information-theoretic framework. These measures show much resemblance to the corresponding entropy measures discussed in the preceeding chapter. The properties of divergence measures and their correlation will also be presented.

2.1 Shannon directed divergence

Let $P = \{p_1,\ldots,p_n\}$ and $Q = \{q_1,\ldots,q_n\}$ denote finite complete discrete probability distributions with $p_i > 0$, $(i=1,\ldots,n)$, $q_j > 0 (j = 1,\ldots,n)$ and

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{j=1}^{n} q_j = 1.$$

The Shannon directed divergence of the distribution Q from the distribution P is given by (Mathai, 1975)

$$J_n\left[\begin{matrix}P\\Q\end{matrix}\right] = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}, \tag{2.1.1}$$

with a convention $0 \log 0 = 0$, to which we shall adhere throughout this section. Analogously, we can introduce the directed divergence of p from Q

$$J_n\left[\begin{matrix}Q\\P\end{matrix}\right] = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}. \tag{2.1.2}$$

Following Kullback and Leibler (1951), we can also introduce a (symmetric) divergence between P and Q

$$\begin{aligned} J_n(P,Q) &= J_n\left[\begin{matrix}P\\Q\end{matrix}\right] + J_n\left[\begin{matrix}Q\\P\end{matrix}\right] \\ &= \sum_{i=1}^{n} (p_i - q_i) \log \frac{p_i}{q_i} = \sum_{i=1}^{n} (q_i - p_i) \log \frac{q_i}{p_i} \end{aligned} \tag{2.1.3}$$

For n = 2, the expression (2.1.1) reduces to

$$J_2\left[\begin{matrix}P\\Q\end{matrix}\right] = J\left[\begin{matrix}p,1-p\\q,1-q\end{matrix}\right] = p \log \frac{p}{q} - (1-p) \log \frac{1-p}{1-q}, \tag{2.1.4}$$

called the directed divergence function, which is an analogue of the entropy function given in (1.1.8).

Suppose P is a prior probability distribution and Q is a posterior distribution corresponding to the state of knowledge after observing one of n possible outcomes in a given experiment. The directed divergence (2.1.1) seems to be a reasonable measure of information obtained through one observation (Rényi, 1961). The basic properties

of the directed divergence (2.1.1) listed below are similar to those of the Shannon entropy.

1. Non-negativity

$$J_n\begin{bmatrix}P\\Q\end{bmatrix} \geq 0, \tag{2.1.5}$$

with equality iff $p_i = q_i$ for all $i = 1,\ldots,n$.

2. Symmetry

$$J_n\begin{bmatrix}p_{k1},\ldots,p_{kn}\\q_{k1},\ldots,q_{kn}\end{bmatrix} = J_n\begin{bmatrix}p_1,\ldots,p_n\\q_1,\ldots,q_n\end{bmatrix}, \tag{2.1.6}$$

where k is an arbitrary permutation on $\{1,\ldots,n\}$ (note that k is the same permutation for both P and Q).

3. Expansibility

$$J_{n+1}\begin{bmatrix}P,0\\Q,0\end{bmatrix} = J_n\begin{bmatrix}P\\Q\end{bmatrix}. \tag{2.1.7}$$

4. Continuity

$J_n\begin{bmatrix}p_1,\ldots,p_n\\q_1,\ldots,q_n\end{bmatrix}$ is a continuous function of all p_i, q_i $(i=1,\ldots,n)$ (2.1.8)

5. Recursivity

$$J_n\begin{bmatrix}p_1,\ldots,p_n\\q_1,\ldots,q_n\end{bmatrix} = J_{n-1}\begin{bmatrix}p_1+p_2,p_3,\ldots,p_n\\q_1+q_2,q_3,\ldots,q_n\end{bmatrix} +$$

$$+(p_1+p_2)\cdot J_2\begin{bmatrix}\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\\\frac{q_1}{q_1+q_2},\frac{q_2}{q_1+q_2}\end{bmatrix},$$

where $p_1 + p_2 > 0$, $q_1 + q_2 > 0$.

6. Strong additivity

Let $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$ $(i=1,\ldots,n;\ j = 1,\ldots,m)$ denote finite complete discrete probability distributions with

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1,$$

$$p_i = \sum_{j=1}^{m} p_{ij} > 0 \text{ and } \sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij} = 1,\quad q_i = \sum_{j=1}^{m} q_{ij} > 0.$$

Then the following equality holds

$$J_{nm}\begin{bmatrix} p_{11}, p_{12}, \ldots, p_{nm} \\ q_{11}, q_{12}, \ldots, q_{nm} \end{bmatrix} = J_n\begin{bmatrix} p_1, \ldots, p_n \\ q_1, \ldots, q_n \end{bmatrix} + \sum_{i=1}^{n} p_i \; J_m\begin{bmatrix} \frac{p_{i1}}{p_i}, \ldots, \frac{p_{im}}{p_i} \\ \frac{q_{i1}}{q_i}, \ldots, \frac{q_{im}}{q_i} \end{bmatrix} \tag{2.1.10}$$

7. Weak additivity

Let $R = \{r_1,\ldots,r_m\}$ and $S = \{s_1,\ldots,s_m\}$ denote complete finite discrete probability distributions with $r_i > 0$ $(i=1,\ldots,m)$, $s_j > 0$ $(j=1,\ldots,m)$ and

$$\sum_{i=1}^{m} r_i = 1, \quad \sum_{j=1}^{m} s_j = 1.$$ Then the following holds

$$J_{nm}\begin{bmatrix} p_1 r_1, \ldots, p_1 r_m, p_2 r_1, \ldots, p_2 r_m, \ldots, p_n r_1, \ldots, p_n r_m \\ q_1 s_1, \ldots, q_1 s_m, q_2 s_1, \ldots, q_2 s_m, \ldots, q_n s_1, \ldots, q_n s_m \end{bmatrix}$$

$$= J_n\begin{bmatrix} p_1, \ldots, p_n \\ q_1, \ldots, q_n \end{bmatrix} + J_m\begin{bmatrix} r_1, \ldots, r_m \\ s_1, \ldots, s_m \end{bmatrix} \tag{2.1.11}$$

A characterization of the directed divergence as defined in (2.1.1) due to Kannapan and Rathie (1973) is based on the following postulates: (2.1.6), for n = 3; (2.1.9),

$$J_2 \begin{bmatrix} \frac{2}{3}, & \frac{1}{3} \\ \frac{1}{3}, & \frac{2}{3} \end{bmatrix} = \frac{1}{3} \tag{2.1.12}$$

and

$$J_2 \begin{bmatrix} p, & 1-p \\ q, & 1-q \end{bmatrix} = 0, \quad p \in (0,1). \tag{2.1.13}$$

Besides, the divergence function J_2 given in (2.1.4) is supposed to have continuous first partial derivatives with respect to both p and q (regularity condition). Another characterization suggested by Kannapan (1972) involves a functional equation corresponding to that given in (1.1.14). Other characterization theorems can be found in (Mathai, 1975).

2.2. Directed divergence of order α

The directed divergence of α (Rényi, 1961) is given by

$${}_{\alpha}J_n \begin{bmatrix} P \\ Q \end{bmatrix} = \frac{1}{\alpha-1} \log \sum_{i=1}^{n} \frac{p_i^{\alpha}}{q_i^{\alpha-1}}, \quad \alpha \neq 1 \tag{2.2.1}$$

When $\alpha \to 1$, (2.2.1) reduces to the Shannon directed divergence as defined in (2.1.1).

The following properties of the Shannon directed divergence also hold for divergence of order α:

1. Non-negativity, as defined in (2.1.5).
2. Symmetry, as defined in (2.1.6).
3. Expansibility, as defined in (2.1.7).
4. Continuity, as defined in (2.1.8).

5. Weak additivity, as defined in (2.1.11).

Analogously to the entropy of order α, the divergence of order α fails to be recursive and strongly additive. An extension of the concept of directed divergence of order α to incomplete probability distributions, as well as its characterization, are given in (Rényi, 1961).

2.3 Directed divergence of type β

A directed divergence associated with the concept of the entropy of type β has been introduced by Rathie and Kannapan (1972). This measure is given by

$$J_n^{\beta}\begin{bmatrix} P \\ Q \end{bmatrix} = \frac{1-\sum_{i=1}^{n} \frac{p_i^{\beta}}{q_i^{\beta-1}}}{1-2^{\beta-1}}, \quad \beta \neq 1, \tag{2.3.1}$$

with a usual convention $0^{\beta} = 0 \quad (\beta \neq 0)$.
When $\beta \to 1$, the divergence of type β reduces to the Shannon divergence defined by (2.1.1). The following equality relating the divergence of type β and the divergence of order α can be derived from (2.2.1) and (2.3.1)

$$J_n^{\beta}\begin{bmatrix} P \\ Q \end{bmatrix} = \frac{1-2^{(\beta-1)\,{}_{\beta}J_n\left[\begin{smallmatrix} P \\ Q \end{smallmatrix}\right]}}{1-2^{\beta-1}}, \quad \beta \neq 1. \tag{2.3.2}$$

For n = 2, (2.3.1) reduces to

$$J_2^{\beta}\begin{bmatrix} p,1-p \\ q,1-q \end{bmatrix} = \frac{1-\frac{p^{\beta}}{q^{\beta-1}}-\frac{(1-p)^{\beta}}{(1-q)^{\beta-1}}}{1-2^{\beta-1}}, \quad \beta \neq 1, \tag{2.3.3}$$

called directed divergence function of type β.

The directed divergence of type β can be expressed in terms of the

directed divergence function by

$$J_n^{\beta}\begin{bmatrix}P\\Q\end{bmatrix} = \sum_{i=2}^{n} \frac{r_i^{\beta}}{s_i^{\beta-1}} f\left(\frac{p_i}{r_i}, \frac{q_i}{s_i}\right), \tag{2.3.4}$$

where

$$f(x,y) = J_2^{\beta}\begin{bmatrix}x, 1-x\\ y, 1-y\end{bmatrix} \tag{2.3.5}$$

is the directed divergence of type β mentioned above and

$$r_i = p_1 + \ldots + p_i,\ s_i = q_1 + \ldots + q_i\ (i = 2,\ldots,n).$$

The main properties of the directed divergence of type β can be seen from the list below.

1. Non-negativity, as defined in (2.1.5).
2. Symmetry, as defined in (2.1.6).
3. Expansibility, as defined in (2.1.7).
4. Recursivity of type β

$$J_n^{\beta}\begin{bmatrix}p_1,\ldots,p_n\\ q_1,\ldots,q_n\end{bmatrix} = J_{n-1}^{\beta}\begin{bmatrix}p_1+p_2, p_3,\ldots,p_n\\ q_1+q_2, q_3,\ldots,q_n\end{bmatrix} + \frac{(p_1+p_2)^{\beta}}{(q_1+q_2)^{\beta-1}} J_2^{\beta}\begin{bmatrix}\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\\ \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\end{bmatrix}$$

for $p_1 + p_2 > 0,\ q_1 + q_2 > 0.$ (2.3.6)

5. Strong additivity of type β

$$J^{\beta}_{nm}\begin{bmatrix} p_{11},p_{12},\cdots,p_{nm} \\ q_{11},q_{12},\cdots,q_{nm} \end{bmatrix} = J^{\beta}_{n}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n \end{bmatrix} + \sum_{i=1}^{n} \frac{p_i^{\beta}}{q_i^{\beta}-1} \cdot J^{\beta}_{m}\begin{bmatrix} \frac{p_{i_1}}{p_i},\cdots,\frac{p_{im}}{p_i} \\ \frac{q_{i_1}}{q_i},\cdots,\frac{q_{im}}{q_i} \end{bmatrix} \quad (2.3.7)$$

with the same notations as in (2.1.10).

6. Weak additivity of type β

$$J^{\beta}_{nm}\begin{bmatrix} p_1r_1,\cdots,p_1r_m,\cdots,p_nr_1,\cdots,p_nr_m \\ q_1s_1,\cdots,q_1s_m,\cdots,q_ns_1,\cdots,q_ns_m \end{bmatrix} = J^{\beta}_{n}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n \end{bmatrix} + J^{\beta}_{m}\begin{bmatrix} r_1,\cdots,r_m \\ s_1,\cdots,s_m \end{bmatrix} - (1-2^{\beta-1})\ J^{\beta}_{n}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n \end{bmatrix} J^{\beta}_{m}\begin{bmatrix} r_1,\cdots,r_m \\ s_1,\cdots,s_m \end{bmatrix} \quad (2.3.8)$$

with the same notations as in (2.1.11).

For characterizations of the directed divergence of type β, see (Rathie, 1972; Mathai, 1975; Patni, 1976).

2.4 Other generalizations of directed divergence

A generalization of the divergence of type β has been introduced by Sharma and Autar (1974). It is defined by

$$J_n^{\alpha,\beta}\left[{P \atop Q}\right] = \frac{1 - \sum_{i=1}^{n} \frac{p_i^{\beta}}{q_i^{\beta-\alpha}}}{1-2^{\beta-\alpha}}, \quad \alpha \neq \beta, \tag{2.4.1}$$

called directed divergence of type (α,β). This measure has been characterized by Patni and Jain (1976).

Several other generalizations can be found in (Kapur, 1968; Rathie, 1971; Mathai, 1975). The concept of divergence can be extended to three or more probability distributions. Let $P = \{p_1,\ldots,p_n\}$, $Q = \{q_1,\ldots q_n\}$ and $R = \{r_1,\ldots,r_n\}$ denote finite complete probability distributions with $p_i \geq 0$, $q_i \geq 0$, $r_i \geq 0$ $(i = 1,\ldots,n)$ and

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} r_i = 1.$$

The directed divergence of R from Q with respect to P is given by (Mathai, 1975)

$$J_n\left[P,{Q \atop R}\right] = \sum_{i=1}^{n} p_i \log \frac{q_i}{r_i}, \tag{2.4.2}$$

with a convention $0 \log 0 = 0$.

A generalized divergence of order α involving three distributions can be defined by

$${}^{\alpha}J_n\left[P,{Q \atop R}\right] = \frac{1}{\alpha-1} \log \sum_{i=1}^{n} p_i \frac{q_i^{\alpha-1}}{r_i^{\alpha-1}}, \quad \alpha \neq 1, \; 0^{\alpha} = 0, \; (\alpha \neq 0) \tag{2.4.3}$$

In a case where $\alpha \to 1$, (2.4.3) reduces to (2.4.2), and with $Q = P$ we obtain (2.2.1).

The generalized divergence of type β for three probability distributions is given by

$$J_n^{\beta}\left[\begin{matrix}P,\\ R\end{matrix}\ \begin{matrix}Q\\ \end{matrix}\right] = \frac{1 - \sum_{i=1}^{n} p_i \frac{q_i^{\beta-1}}{r_i^{\beta-1}}}{1 - 2^{\beta-1}}, \quad \beta \neq 1, \quad 0^{\beta} = 0 \ (\beta \neq 0) \tag{2.4.4}$$

Again, with $\beta \to 1$ (2.4.4) reduces to (2.4.2), and for $Q = P$ we have the directed divergence of type β as defined in (2.3.1). The generalized divergences of order α and of type β are related by

$$J_n^{\beta}\left[\begin{matrix}P,\\ R\end{matrix}\ \begin{matrix}Q\\ \end{matrix}\right] = \frac{1-2^{(\beta-1)}{}_{\beta}J_n(P,{}^{Q}_{R})}{1 - 2^{\beta-1}}, \quad \beta \neq 1, \tag{2.4.5}$$

which can be obtained from (2.4.3) and (2.4.4).

The properties of divergences involving three distributions are analogous to the corresponding properties, which hold for divergence measures discussed in the preceeding sections. The recursivity property of the generalized divergence of type β, for instance, can be expressed as follows:

$$J_n^{\beta}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n \\ r_1,\cdots,r_n \end{bmatrix} = J_{n-1}^{\beta}\begin{bmatrix} p_1+p_2,p_3,\cdots,p_n \\ q_1+q_2,q_3,\cdots,q_n \\ r_1+r_2,r_3,\cdots,r_n \end{bmatrix} +$$

$$+ (p_1+p_2)\ \frac{(q_1+q_2)^{\beta-1}}{(r_1+r_2)^{\beta-1}}\ J_2^{\beta}\begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{q_1}{q_1+q_2} & \frac{q_2}{q_1+q_2} \\ \frac{r_1}{r_1+r_2} & \frac{r_2}{r_1+r_2} \end{bmatrix} \tag{2.4.6}$$

The expression for the weak additivity of type β attains the form

$$J_{nm}^{\beta}\begin{bmatrix} p_1s_1,\ldots,p_1s_m,\ldots,p_ns_1,\ldots,p_ns_m \\ q_1t_1,\ldots,q_1t_m,\ldots,q_nt_1,\ldots,q_nt_m \\ r_1u_1,\ldots,r_1u_m,\ldots,r_nu_1,\ldots,r_nu_m \end{bmatrix} = J_n^{\beta}\begin{bmatrix} p_1,\ldots,p_n \\ q_1,\ldots,q_n \\ r_1,\ldots,r_n \end{bmatrix} +$$

$$+ J_m^{\beta}\begin{bmatrix} s_1,\ldots,s_m \\ t_1,\ldots,t_m \\ u_1,\ldots,u_m \end{bmatrix} - (1-2^{\beta-1})\ J_n^{\beta}\begin{bmatrix} p_1,\ldots,p_n \\ q_1,\ldots,q_n \\ r_1,\ldots,r_n \end{bmatrix} J_m^{\beta}\begin{bmatrix} s_1,\ldots,s_m \\ t_1,\ldots,t_m \\ u_1,\ldots,u_m \end{bmatrix} \qquad (2.4.7)$$

with $p_i > 0$, $q_i > 0$, $r_i > 0$, $s_j > 0$, $t_j > 0$, $u_j > 0$ $(i=1,\ldots,n;\ j = 1,\ldots,m)$ and

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} r_i = \sum_{j=1}^{m} s_j = \sum_{j=1}^{m} t_j = \sum_{j=1}^{m} u_j = 1.$$

(see Mathai, 1975).

3. MEASURES OF INACCURACY

This chapter deals with the concept of inaccuracy introduced by Kerridge (1961). Inaccuracy can be considered as a generalization of both entropy and divergence. We shall show that the measures of inaccuracy have many properties of the corresponding measures of entropy and divergence. The relations of inaccuracy measures to the measures of entropy and divergence are also discussed.

3.1 Shannon inaccuracy

Consider two finite complete discrete probability distributions

$P = \{p_1,\ldots,p_n\}$ and $Q = \{q_1,\ldots,q_n\}$ with $p_i > 0$, $q_i > 0$ $(i=1,\ldots,n)$

and

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1.$$

Inaccuracy (of Q with respect to P) can be defined by

$$\bar{A}_n\begin{bmatrix} P \\ Q \end{bmatrix} = -\sum_{i=1}^{n} p_i \log q_i, \qquad (3.1.1)$$

with a usual convention 0 log 0 = 0. It is also supposed that $q_i = 0$ implies $p_i = 0$, for all $i = 1,\ldots,n$.
Definition given in (3.1.1) due to Kerridge (1961) is similar to the definition of Shannon's entropy as given in (1.1.1). It also resembles the definition of Shannon's directed divergence (2.1.1). In fact, the following relation holds

$$\bar{A}_n \begin{bmatrix} P \\ Q \end{bmatrix} = H_n(P) + J_n \begin{bmatrix} P \\ Q \end{bmatrix}, \qquad (3.1.2)$$

which can be easily obtained on the account of (1.1.1), (2.1.1) and (3.1.1). The equality (3.1.2) shows that Shannon's inaccuracy of one distribution with respect to another distribution is a sum of the Shannon entropy for the first distribution and the Shannon divergence of the second distribution from the first one.

Suppose P is a true distribution due to the instrinsic randomness of a certain phenomenon and Q is an estimate of P based on (possibly incorrect) knowledge available. In this case inaccuracy (3.1.2) can be considered as a measure of total uncertainty concerning the phenomenon in question, which is due to both its intrinsic vagueness (given by P) and to inaccuracy of the knowledge about this vagueness (given by Q).

The properties of Shannon's inaccuracy listed below resemble, in many respects, those of Shannon's directed divergence and entropy, which can be expected on account of the relation (3.1.2).

1. Non-negativity, as given in (2.1.5), with equality iff $p_i = q_i = 1$ for some i (i=1,...,n). (3.1.3)

2. Symmetry, as defined in (2.1.6).
3. Expansibility, as defined in (2.1.7).
4. Continuity, as defined in (2.1.8).
5. Recursivity, as defined in (2.1.9).
6. Strong additivity, as defined in (2.1.10).

7. Weak additivity, as defined in (2.1.11).

8. Monotony

$$\bar{A}_n \begin{bmatrix} \frac{1}{n},\cdots,\frac{1}{n} \\ \frac{1}{n},\cdots,\frac{1}{n} \end{bmatrix} \leq \bar{A}_{n+1} \begin{bmatrix} \frac{1}{n+1},\cdots,\frac{1}{n+1} \\ \frac{1}{n+1},\cdots,\frac{1}{n+1} \end{bmatrix} \qquad (3.1.4)$$

(cf. (1.1.24)).

9. Minimality

$$\inf_{Q} \bar{A}_n \begin{bmatrix} P \\ Q \end{bmatrix} = \bar{A}_n \begin{bmatrix} P \\ P \end{bmatrix} = H_n(P), \qquad (3.1.5)$$

where $H_n(P)$ is the Shannon entropy as defined in (1.1.1) (cf. (1.1.26)).

10. Maximality

$$\sup_{P} \inf_{Q} \bar{A}_n \begin{bmatrix} P \\ Q \end{bmatrix} = \bar{A}_n \begin{bmatrix} \frac{1}{n},\cdots,\frac{1}{n} \\ \frac{1}{n},\cdots,\frac{1}{n} \end{bmatrix} \qquad (3.1.6)$$

The latter two properties, which follow from (1.1.1), (2.1.1) and (3.1.1) on account of (3.1.2), are typical of Shannon's inaccuracy.

For n = 2, (3.1.1) attains the form

$$f(p,q) = \bar{A}_2 \begin{bmatrix} p,1-p \\ q,1-q \end{bmatrix} = -p \log p - (1-p) \log (1-q), \qquad (3.1.7)$$

called inaccuracy function, which is an aanalogue of the entropy function as defined in (1.1.8). This measure can be shown to be a solution of a functional equation given by (Kannapan, 1972)

$$f(x,y) + (1-x)f\left(\frac{u}{1-x},\frac{v}{1-y}\right) = f(u,v) + (1-u)f\left(\frac{x}{1-u},\frac{y}{1-v}\right) \qquad (3.1.8)$$

with $x,y,u,v \in [0,1)$ and $x+u,\ y+v \in [0,1]$. Equation (3.1.8) is a generalization of the functional equation given in (1.1.14).

An expression for the Shannon inaccuracy in terms of the function (3.1.7) is given by

$$\bar{A}_n\left[\begin{matrix}P\\Q\end{matrix}\right] = \sum_{i=2}^{n} r_i \bar{A}_2 \left[\frac{r_{i-1}}{r_i}, \frac{p_i}{r_i}; \frac{s_{i-1}}{s_i}, \frac{q_i}{s_i}\right], \qquad (3.1.9)$$

where $r_i = p_1+\ldots+p_i$, $s_i = q_1+\ldots q_i$ $(i=2,\ldots,n)$.

It can be easily derived on account of (3.1.1) and (3.1.7). Several characterizations of the Shannon inaccuracy are known (Kerridge, 1961; Rathie, 1971; Kannapan, 1972; Mathai, 1975).

3.2 Inaccuracy of order α

A generalization of Shannon's inaccuracy corresponding to both the entropy of order α and the directed divergence of order α can be defined by

$$\alpha\bar{A}_n\left[\begin{matrix}P\\Q\end{matrix}\right] = {}_\alpha H_n(P) + {}_\alpha J_n\left[\begin{matrix}P\\Q\end{matrix}\right] =$$

$$= \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_i\, q_i^{\alpha-1}, \quad \alpha \neq 1, \qquad (3.2.1)$$

with ${}_\alpha H_n(P)$ and ${}_\alpha J_n\left[\begin{matrix}P\\Q\end{matrix}\right]$ as defined in (1.2.10) and (2.2.1) respectively. In this definition the usual convention $0^\alpha=0$ $(\alpha\neq 0)$ is followed.

The properties of this measure are similar to those of ${}_\alpha H_n(P)$ and of ${}_\alpha J_n\left[\begin{matrix}P\\Q\end{matrix}\right]$. Some of them are listed below.

1. Non-negativity, as defined in (3.1.3).
2. Symmetry, as defined in (2.1.6).
3. Expansibility, as defined in (2.1.7).
4. Continuity, as defined in (2.1.8).

5. Weak additivity, as defined in (2.1.11).
6. Monotony, as defined in (3.1.4).
7. Minimality, as defined in (3.1.5) (after replacing $\bar{A}_n\left[\begin{smallmatrix}P\\Q\end{smallmatrix}\right]$ by ${}_\alpha A_n\left[\begin{smallmatrix}P\\Q\end{smallmatrix}\right]$ and $H_n(P)$ by ${}_\alpha H_n(P)$).
8. Maximality, as defined by (3.1.6).

The properties 5 and 6 of the Shannon inaccuracy (recursivity and strong additivity) are not satisfied by the inaccuracy of order α.

3.3 Inaccuracy of type β

An alternative generalization of Shannon's inaccuracy is given by

$$\bar{A}_n^\beta\left[\begin{smallmatrix}P\\Q\end{smallmatrix}\right] = \frac{1-\sum_{i=1}^{n} p_i\, q_i^{\beta-1}}{1-2^{1-\beta}}, \quad \beta \neq 1, \tag{3.3.1}$$

called inaccuracy of type β (Mathai, 1975). In contrast to (3.2.1) this definition does not admit a representation of inaccuracy by a sum of the corresponding measures of entropy and directed divergence.

Another form of (3.3.1) can be obtained by making use of a so-called inaccuracy function of type β given by

$$f^\beta(p,q) = \underset{\beta\neq 1}{\bar{A}_2^\beta}\left[\begin{smallmatrix}P\\Q\end{smallmatrix}\right] = A_2^\beta\left[\begin{smallmatrix}p,1-p\\q,1-q\end{smallmatrix}\right] = \frac{1-pq^{\beta-1}-(1-p)(1-q)^{\beta-1}}{1-2^{1-\beta}} \tag{3.3.2}$$

being a solution of a functional equation

$$f(x,y) + (1-x)(1-y)^{\beta-1} f\left(\frac{u}{1-x},\frac{v}{1-y}\right) = f(u,v) + (1-u)(1-v)^{\beta-1} f\left(\frac{x}{1-u},\frac{y}{1-v}\right), \tag{3.3.3}$$

with $x,y,u,v \in [0,1)$; $x+u,\ y+v \in [0,1]$, under the following boundary conditions

$$f(0,0) = f(1,1) \tag{3.3.4}$$
$$f(\tfrac{1}{2},\tfrac{1}{2}) = 1 \tag{3.3.5}$$

On account of (3.3.2), another expression for the inaccuracy of type β is given by (Mathai, 1975)

$$\bar{A}_n^{\beta}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n \end{bmatrix} = \sum_{i=2}^{n} r_i s_i^{\beta-1} f^{\beta}\left(\frac{p_i}{r_i}, \frac{q_i}{s_i}\right), \tag{3.3.6}$$

where $r_i = p_1,\cdots,p_i$, $s_i = q_1,\cdots,q_i$ $(i=2,\cdots,n)$.

Next we give the basic properties of the inaccuracy of type β, which show much resemblance to those of the directed divergence of type β.

1. Non-negativity, as given in (3.1.3).
2. Symmetry, as defined in (2.1.6).
3. Expansibility, as defined in (2.1.7).
4. Continuity, as defined in (2.1.8).
5. Recursivity of type β.

$$\bar{A}_n^{\beta}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n \end{bmatrix} = \bar{A}_{n-1}^{\beta}\begin{bmatrix} p_1+p_2,p_3,\cdots,p_n \\ q_1+q_2,q_3,\cdots,q_n \end{bmatrix} +$$

$$+ (p_1+p_2)(q_1+q_2)^{\beta-1} A_2^{\beta}\begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{q_1}{q_1+q_2} & \frac{q_2}{q_1+q_2} \end{bmatrix}, \tag{3.3.7}$$

where $p_1 + p_2 > 0$, $q_1 + q_2 > 0$.

6. Strong additivity of type β

$$\bar{A}^{\beta}_{nm}\begin{bmatrix} p_{11},p_{12},\cdots,p_{nm} \\ q_{11},q_{12},\cdots,q_{nm}\end{bmatrix} = \bar{A}^{\beta}_{n}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n\end{bmatrix} + \sum_{i=1}^{n} p_i q_i^{\beta-1} \cdot \bar{A}^{\beta}_{m}\begin{bmatrix} \frac{p_{i1}}{p_i},\cdots,\frac{p_{im}}{p_i} \\ \frac{q_{i1}}{q_i},\cdots,\frac{q_{im}}{q_i}\end{bmatrix}, \quad (3.3.8)$$

with the same notations as in (2.1.10).

7. Weak additivity of type β

$$\bar{A}^{\beta}_{nm}\begin{bmatrix} p_1r_1,\cdots,p_1r_m,\cdots,p_nr_1,\cdots p_nr_m \\ q_1s_1,\cdots,q_1s_m,\cdots q_ns_1,\cdots,q_ns_m\end{bmatrix} = \bar{A}^{\beta}_{n}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n\end{bmatrix} + \bar{A}^{\beta}_{m}\begin{bmatrix} r_1,\cdots,r_m \\ s_1,\cdots,s_m\end{bmatrix} + (1-2^{1-\beta})\ \bar{A}^{\beta}_{n}\begin{bmatrix} p_1,\cdots,p_n \\ q_1,\cdots,q_n\end{bmatrix}\ \bar{A}^{\beta}_{m}\begin{bmatrix} r_1,\cdots,r_m \\ s_1,\cdots,s_m\end{bmatrix}, \quad (3.3.9)$$

with the same notations as in (2.1.11).

Note that the properties (3.3.7) to (3.3.9) do not coincide with the corresponding properties of the directed divergence of type β given in (2.3.6) to (2.3.8).

A characterization theorem due to Rathie and Kannapan (1973) involves the following postulates: recursivity, as defined in (3.3.7); symmetry, as defined in (2.1.6) and (a boundary condition)

$$\bar{A}^{\beta}_{2}\begin{bmatrix} \frac{1}{2},\frac{1}{2} \\ \frac{1}{2},\frac{1}{2}\end{bmatrix} = 1. \quad (3.3.10)$$

3.4 Inaccuracy of type (β,γ)

A further generalisation of the Shannon inaccuracy measure, called inaccuracy of type (β,γ), has been introduced by Sharma and Autar (1973). The inaccuracy of type (β,γ) is given by a sum

$$\bar{A}_n^{\beta\gamma}\begin{bmatrix}P\\Q\end{bmatrix} = \sum_{i=2}^{n} r_i^{\gamma}\, s_i^{\beta-\gamma}\; f^{\beta,\gamma}\left(\frac{p_i}{r_i}\;\frac{q_i}{s_i}\right), \tag{3.4.1}$$

where $r_i = p_1 + \cdots + p_i$, $s_i = q_1 + \cdots + q_i$ $(i=2,\ldots,n)$ and $f^{\beta,\gamma}$ is the inaccuracy function of type (β,γ) defined by

$$f^{\beta,\gamma}(p,q) = \bar{A}_2^{\beta,\gamma}\begin{bmatrix}p,1-p\\q,1-q\end{bmatrix} = \frac{1-p^{\gamma}q^{\beta-\gamma}-(1-p)^{\gamma}(1-q)^{\beta-\gamma}}{1-2^{1-\beta}}, \tag{3.4.2}$$

$\beta,\gamma > 0$; $\beta \neq 1$ when $\gamma = 1$.

Setting (3.4.2) in (3.4.1) results in another expression for the inaccuracy measure of type (β,γ)

$$\bar{A}_n^{\beta,\gamma}\begin{bmatrix}P\\Q\end{bmatrix} = \frac{1-\sum_{i=1}^{n} p_i^{\gamma}\, q_i^{\beta-\gamma}}{1-2^{1-\beta}}, \tag{3.4.3}$$

The function of $f^{\beta,\gamma}$ appears to be a solution of a functional equation

$$f(x_1,y_1) + (1-x_1)^{\gamma}(1-y_1)^{\beta-\gamma}\, f\left(\frac{x_2}{1-x_1}, \frac{y_2}{1-y_1}\right) =$$

$$= f(x_2,y_2) + (1-x_2)^{\gamma}(1-y_2)^{\beta-\gamma}\, f\left(\frac{x_1}{1-x_2}, \frac{y_1}{1-y_2}\right) \tag{3.4.4}$$

under the following conditions

$$f(1,1) = f(0,0), \tag{3.4.5}$$

$$f(\tfrac{1}{2},\tfrac{1}{2}) = 1. \tag{3.4.6}$$

For $\gamma = 1$, the inaccuracy of type (β,γ) reduces to the inaccuracy of type β, as defined in (3.3.1), and for $P = Q$ it becomes an entropy of type β given by (1.3.7), with any γ. A characterization theorem due to Sharma and Autar (1973) is based on the following postulates:

1. Symmetry, as defined in (2.1.6), for n = 3. (3.4.7)

2. Recursivity of type (β,γ)

$$\bar{A}_n^{\beta,\gamma}\begin{bmatrix} p_1,\ldots,p_n \\ q_1,\ldots,q_n \end{bmatrix} = \bar{A}_{n-1}^{\beta,\gamma}\begin{bmatrix} p_1+p_2,p_3,\ldots,p_n \\ q_1+q_2,q_3,\ldots,q_n \end{bmatrix} + \frac{(p_1+p_2)^{\gamma}}{(q_1+q_2)^{\gamma-\beta}} \cdot A_2^{\beta,\gamma}\begin{bmatrix} \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} \\ \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2} \end{bmatrix} \tag{3.4.8}$$

3. Normality

$$\bar{A}_2^{\beta,\gamma}\begin{bmatrix} \frac{1}{2}, \frac{1}{2} \\ \frac{1}{2}, \frac{1}{2} \end{bmatrix} = 1 \tag{3.4.9}$$

Other properties of the inaccuracy of type (β,γ) can be derived as consequences of those mentioned in (3.4.7) to (3.4.9). Some different generalizations of the inaccuracy measure can be found in (Rathie, 1970; 1971; 1972; Mathai, 1975).

4. MEASURES OF CERTAINTY

The concept of certainty appears already in the foundations of the theory of probability. Aczél and Daroczy (1975) have introduced a measure of an average probability, which is, in fact, another term for certainty. This concept proved to be useful for generalizations of Shannon's entropy. An explicit definition of the concept of certainty, as well as several measures of certainty is accredited to Van der Lubbe (1981), whose approach will be followed in this chapter.
The certainty measures involving one or two discrete probability

distributions, their properties and relations between them, will be discussed.

4.1 Marginal measures of certainty

Let $P = \{p_1,\ldots,p_n\}$ and $Q = \{q_1,\ldots,q_m\}$ denote finite complete probability distributions with $p_i > 0$, $(i=1,\ldots,n)$, $q_j > 0$ $(j=1,\ldots,m)$,

$\sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_j = 1$ and let $r = \{r_{ij}\}$ $(i=1,\ldots,n;\ j=1,\ldots,m)$ be another finite complete distribution related to both P and Q, with $\sum_{j=1}^{m} r_{ij} = p_i > 0$ $(i=1,\ldots,n)$, $\sum_{i=1}^{n} r_{ij} = q_j > 0$ $(j=1,\ldots,m)$, $r_{ij} > 0$ $(i=1,\ldots,n;\ j=1,\ldots,m)$, $\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} = 1$. We shall also use a notation q_{ji} for $\frac{r_{ij}}{p_i}$ and p_{ij} for $\frac{r_{ij}}{q_j}$ $(i=1,\ldots,n;\ j=1,\ldots,m)$.

The concept of certainty concerning a random event A can be explicated by making use of its probability p(A). Following Van der Lubbe, we can consider the certainty as a monotonically increasing function of the probability

$$c(A) = f\,(p(A)). \qquad (4.1.1)$$

Such an approach seems to be consistent with the intuitive idea of certainty. Surely, an increase in probability should imply a corresponding increase in certainty and vice versa. In order to further specify the function f, let us consider another event, say B, whose probability is p(B). Suppose B is independent of A and consider a compound event (A,B) comprizing both A and B. It also seems reasonable to require that the function f be such that the degree of certainty associated with (A,B) would be equal to the product of the

degrees of certainty concerning each single event, i.e.

$$c\,(p(A,B)) = c\,(p(A)p(B)) = c\,(p(A))\;c\,(P(B)). \qquad (4.1.2)$$

Furthermore, we may expect f to be a continuous function of the probability with $[0,1]$ as a definition domain. It can be proved that these requirements result in the following general solution of the functional equation (4.1.2)

$$c(p) = p^{a}, \qquad a > 0, \qquad (4.1.3)$$

which can be regarded as a definition of a certainty associated with a single event, called a <u>self-certainty</u>. Actually, self-certainty is a measure of certainty concerning the occurence of a single event.

Let us introduce a concept of a <u>certainty function</u> C(A), which should give a measure of certainty concerning both the occurence and non-occurence of a single event. To obtain a reasonable definition of C(A), we introduce a concept of relative certainty.
Following the argument presented in section 1.1, we consider two disjoint events A and B whose probabilities p(A) and p(B) satisfy the following conditions: $p(A) > 0$, $p(B) > 0$, $p(A) + p(B) < 1$. Let A,B and $\bar{A}$, $\bar{B}$ denote the occurence and non-occurence of the events. The conditional probability of A with respect to $\bar{B}$ and the conditional probability of B with respect to $\bar{A}$ are given by

$$p(A/\bar{B}) = \frac{p(A)}{p(\bar{B})} = \frac{p(A)}{1-p(B)} \qquad (4.1.4)$$

and

$$p(B/\bar{A}) = \frac{p(B)}{p(\bar{A})} = \frac{p(B)}{1-p(A)} \qquad (4.1.5)$$

correspondingly.
Next we introduce the concept of a <u>joint certainty</u> concerning two events, C(A,B) and the concept of a <u>relative certainty</u> of one event with respect to another event, denoted by $C(A/\bar{B})$ and $C(B/\bar{A})$.

Suppose C(A,B) is a symmetric function of A and B, satisfying

$$C(A,B) = C(A)\;C(B/\bar{A}) = C(B)\;C(B/\bar{A}). \qquad (4.1.6)$$

Let us require that the certainty function C(A) only depends on the probability p(A),

$$C(A) = f\,(p(A)),\quad C(B) = f\,(p(B)), \tag{4.1.7}$$

and the relative certainty is defined by

$$\begin{aligned} C(A/\bar{B}) &= p(\bar{B})\ f\ (p(A/\bar{B}),\\ C(B/\bar{A}) &= p(\bar{A})\ f\ (p(B/\bar{A}). \end{aligned} \tag{4.1.8}$$

Substituting (4.1.7) and (4.1.8) in (4.1.6) results in

$$f(p(A))\ p(\bar{A})\ f(p(B/\bar{A})) = f(p(B))\ p(\bar{B})\ f(p(A/\bar{B})). \tag{4.1.9}$$

Setting $x = p(A) \in [0,1)$, $y=p(B) \in [0,1)$ in (4.1.9) and taking into consideration (4.1.4), (4.1.5), one gets

$$f(x)\ (1-x)\ f(\frac{y}{1-x}) = f(y)\ (1-y)\ f(\frac{x}{1-y}),\ x + y \leq 1, \tag{4.1.10}$$

which will be termed a <u>functional equation of certainty</u> (cf. (1.1.14). Imposing upon the boundary condition

$$f(1) = f(0), \tag{4.1.11}$$

implying that the certainty about a sure event is equal to the certainty about an impossible event, leads to the following solution of (4.1.10)

$$f(p) = C_2(p,1-p) = [p(1-p)]^a,\quad a > 0. \tag{4.1.12}$$

Note that the certainty function can be considered as a self-certainty of an event $(A,\bar{A})$, consisting of occurence and non-occurence of A. Consider a set of n events with a probability distribution P. There are at least two ways to obtain a measure of certainty about the whole set. We can first determine an average probability for the distribution P and then apply the measure of self-certainty to an "averaged" event whose probability is equal to the average probability of P. But we may also apply the measure of self-certainty to each single event and afterwards determine the certainty of the whole set as an average self-certainty.

Let us take the first way. Analogously to the argument followed in 1.2, we define the <u>average probability</u> of a given distribution P by

$$\overset{\vee}{p}_n \quad \phi^{-1}\left[\sum_{i=1}^{n} p_i \, \phi(p_i)\right], \tag{4.1.13}$$

where ϕ is such a strictly monotonic function of $x \in (0,1]$ that the function

$$\phi^*(x) = \begin{cases} x\,\phi(x) & ; \quad x \in (0,1] \\ 0 & ; \quad x = 0 \end{cases} \tag{4.1.14}$$

be continuous. As has been mentioned in section 1.2, in this case $\overset{\vee}{p}_n$ appears to be symmetric (as defined in (1.2.3)), expansible (as defined in (1.2.4) and it takes on values in $[p_{min}, p_{max}]$, where p_{min} and p_{max} denote the minimal and maximal probabilities in P (cf. 1.2.5). Imposing another reasonable requirement given by

$$\overset{\vee}{p}(P,Q) = \overset{\vee}{p}\,\overset{\vee}{q}, \tag{4.1.15}$$

which implies that the average probability of two independent distributions be equal to the product of the average probabilities of each distribution, we arrive at two possible solutions for $\phi(x)$ (Daróczy, 1964),

$$\phi_1(x) = \log x, \quad x \in (0,1] \tag{4.1.16}$$

and

$$\phi_2(x) = x^{\alpha-1}, \quad x \in (0,1], \ \alpha > 0, \ \alpha \neq 1. \tag{4.1.17}$$

Substituting (4.1.16) and (4.1.17) in (4.1.13) results in

$$\overset{\vee}{p}_n(1) = \prod_{i=1}^{n} p_i^{p_i} \tag{4.1.18}$$

and

$$\overset{\vee}{p}_n(2) = \left(\sum_{i=1}^{n} p_i^{\alpha}\right)^{\frac{1}{\alpha-1}}, \quad \alpha > 0, \quad \alpha \neq 1, \tag{4.1.19}$$

respectively. A measure of certainty associated with a given distribution P can be defined as a self-certainty of a single event whose probability is equal to $\overset{\vee}{p}_n$. Setting $\overset{\vee}{p}_n(1)$ and $\overset{\vee}{p}_n(2)$ in (4.1.3), we obtain two possible measures given by

$$C_n^{(1)}(P) = c(\overset{\vee}{p}_n^{(1)}) = \prod_{i=1}^{n} p_i^{ap_i}, \quad a > 0 \qquad (4.1.20)$$

and

$$C_n^{(2)}(P) = c(p_n^{(2)}) = \left(\sum_{i=1}^{n} p_i^{\alpha} \right)^{\frac{a}{\alpha-1}}, \quad a > 0, \ \alpha > 0, \ \alpha \neq 1. \qquad (4.1.21)$$

With a notation $b = \frac{\alpha-1}{a}$, the latter expression transforms to

$$C_n^{(2)}(P) = \left(\sum_{i=1}^{n} p_i^{ab+1} \right)^{\frac{1}{b}}, \quad a > 0, \ (ab+1) > 0, \ b \neq 0. \qquad (4.1.22)$$

Note that in a limiting case, when $b = 0$, the measure given in (4.1.22) reduces to that given by (4.1.20). We also mention that (4.1.22) is identical to a measure given by (2.2.8) in (Van der Lubbe, 1981), which has been obtained under a somewhat involved requirement that $C_n(P)$ be strictly Shur convex. On account of (4.1.2) and (4.1.15) we have

$$C_{nm}^{(k)}(P,Q) = C_n^{(k)}(P) \ C_m^{(k)}(Q), \ k = 1,2, \qquad (4.1.23)$$

which expresses the so-called multiplicativity property occuring in (Van der Lubbe, 1981) as another requirement imposed upon $C_n(P)$.

In what follows, we mention the basic properties of $C_n^{(2)}(P)$ as defined in (4.1.22). The upper index "(2)" in the latter notation will be omitted throughout the remainder of this section. Wherever pertinent, a reference will be made to the corresponding properties of uncertainty measures discussed in chapter 1. For proofs, we refer to (Van der Lubbe, 1981).

1. Non-negativity (cf. (1.1.19))
   $C_n(P) > 0$ (4.1.24)
2. Symmetry, as defined in (1.1.20).
3. Expansibility, as defined in (1.1.21).
4. Continuity, as defined in (1.1.23).
5. Strict monotony (cf. (1.1.24))

$$C_{n+1}(\frac{1}{n+1},\ldots,\frac{1}{n+1}) < C_n(\frac{1}{n},\ldots,\frac{1}{n}). \qquad (4.1.25)$$

6. Maximality (cf. (1.1.25))

$$C_n(p_1,\ldots,p_n) \leq C_n(0,\ldots,0,1,0,\ldots,0) = 1 \qquad (4.1.26)$$

7. Minimality (cf. (1.1.19), (1.1.32))

$$C_n(P) \geq C_n(\frac{1}{n},\ldots,\frac{1}{n}) = \frac{1}{n^a}. \qquad (4.1.27)$$

8. Weak multiplicativity (cf. (1.1.31))

$$C_{nm}(P,Q) = C_n(P) \quad C_m(Q) \qquad (4.1.28)$$

where P and Q are independent distributions, as defined in the beginning of this section.

9. Inequalities (cf. (1.1.30))

$C_{nm}(R) \leq C_n(P)$, with equality iff R reduces to P;

$C_{nm}(R) \leq C_m(Q)$, with equality iff R reduces to Q, (4.1.29)

where P,Q,R are probability distributions as defined in the beginning of this section.

10. $C_2(p,1-p)$ is a monotonically non-increasing function of p. (4.1.30)

11. For $a > 1$, the following inequality holds

$$p_{min}^a \leq C_n(p_1,\ldots,p_n) \leq p_{max}^a \leq p_{max}, \qquad (4.1.31)$$

where

$p_{min} = \min_i (p_1,\ldots,p_i,\ldots,p_n)$ and $p_{max} = \max_i (p_1,\ldots,p_i,\ldots,p_n)$;

for $(ab+1) > 1$, $\frac{1}{b} > 0$, $0 < a \leq 1$ as well as for

$0 < (ab + 1) < 1$, $\frac{1}{b} < 0$, $0 < a \leq 1$, it holds

$$p_{min} \leq p_{min}^a \leq C_n(P) \leq p_{max}^a. \qquad (4.1.32)$$

12. $C_n(p,\ldots,p_n) \leq C_2(p_i, 1-p_i) = (p_i^{ab+1}+(1-p_i)^{ab+1})^{\frac{1}{b}}$
for all $i = 1,\ldots,n.$ (4.1.33)

13. Convexity
$C_n(P)$ is convex in P for $0 < (ab+1) < 1$, $\frac{1}{b} < 0$ and for
$(ab+1) > 1$, $(a + \frac{1}{b}) > 1$ (4.1.34)

14. Strict Shur convexity (cf. (1.1.26))
$C_n(S) \leq C_n(P)$, (4.1.35)
where $P = \{p_1,\ldots,p_n\}$, $S = \{s_1,\ldots,s_n\}$, $p_i > 0$, $s_i > 0$ $(i=1,\ldots,n)$,

$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} s_i = 1$ and S is an arbitrary Shur transformation of P, i.e. such that $s_j = \sum_{i=1}^{n} t_{ij}p_i$ $(j=1,\ldots,n)$, $t_{ij} > 0$,

$\sum_{i=1}^{n} t_{ij} = \sum_{j=1}^{n} t_{ij} = 1$,

with equality in (4.1.35) iff S is a permutation on P.

The properties given in (4.1.25) to (4.1.27) and (4.1.29) are direct consequences of the strict Shur convexity defined in (4.1.35). It can be seen from (4.1.27) that the constant a determines the lower bound on a measurement scale for certainty. The references made in the list above indicate that much similarity exists between the measures of certainty and those of uncertainty.

We also notice that properties 1 to 14 hold for $C_n^{(1)}(P)$ as well (in 11 to 13, the conditions involving the constant b should be omitted of course).

4.2 Conditional and joint measures of certainty

Let $C_{n/m}(P/Q)$ be a certainty associated with the same probability distribution P, given another distribution Q. The knowledge of Q may

increase the degree of certainty about P, provided P is dependent on Q. For independent P and Q, this knowledge does not contribute to our certainty about P. It seems to be reasonable to require that for any P and Q the following inequality holds

$$C_{n/m}(P/Q) \geq C_n(P), \tag{4.2.1}$$

with equality iff P and Q are independent (Van der Lubbe, 1981).

A measure for $C_{n/m}(P/Q)$ can be derived through an appropriate generalization of the marginal certainty measure. Different generalizations will yield different measures. Replacing marginal probabilities p_i by conditional probabilities p_{ij} in (4.1.20) and in (4.1.22), we obtain conditional certainty measures with respect to a j-th event of Q,

$$C^{(1)}_{n/j}(P/j) = \prod_{i=1}^{n} p_{ij}^{ap_{ij}}, \quad a > 0 \tag{4.2.2}$$

and

$$C^{(2)}_{n/j}(P/j) = \left(\sum_{i=1}^{n} p_{ij}^{ab+1} \right)^{\frac{1}{b}}, \; a > 0, \; ab + 1 > 0, \; b \neq 0, \tag{4.2.3}$$

respectively. Taking mathematical expectation with respect to Q on both sides of the latter expression results in

$$C^{(2)}_{n/m}(P/Q) = \sum_{j=1}^{m} q_j \left(\sum_{i=1}^{n} p_{ij}^{ab+1} \right)^{\frac{1}{b}}, \; a > 0, \; ab+1 > 0, \; b \neq 0, \tag{4.2.4}$$

which is the conditional certainty measure sought.
To obtain a conditional certainty measure corresponding to $C_n^{(1)}(P)$, we should take the expectation of (4.2.2), but this procedure leads to a rather cumbersome expression involving addition with respect to i and multiplication with respect to j. Another possible approach arises by introducing a concept of an average conditional probability, which seems to be of interest as such. Replacing marginal probabilities p_i in (4.1.13) by conditional probabilities p_{ij} and introducing an operation of mathematical expectation with respect to Q prior to ϕ^{-1}, we obtain,

$$\overset{v}{p}_{n/m}(Q) = \phi^{-1}\left[\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{ij}\,\phi(p_{ij})\right], \qquad (4.2.5)$$

which can be regarded as a definition of an average probability of P given Q. Making use of the function ϕ_1 given in (4.1.16) results in

$$\overset{v}{p}_{n/m}(Q) = \prod_{i=1}^{n}\prod_{j=1}^{m} p_{ij}^{q_j p_{ij}}. \qquad (4.2.6)$$

Setting the right hand side of (4.2.6) for p in (4.1.3), one gets a measure of conditional certainty

$$C^{(1)}_{n/m}(P/Q) = \prod_{i=1}^{n}\prod_{j=1}^{m} p_{ij}^{a q_j p_{ij}}, \quad a > 0. \qquad (4.2.7)$$

An alternative definition of a conditional certainty corresponding to the marginal certainty $C_n^{(2)}(P)$ arises if we reverse the order of taking the expectation and raising to the power 1/b on the right hand side of (4.2.4):

$$C^{(2a)}_{n/m}(P/Q) = \left(\sum_{j=1}^{m} q_j \sum_{i=1}^{n} p_{ij}^{ab+1}\right)^{\frac{1}{b}}, \quad a > 0, \qquad (4.2.8)$$

$$(ab+1) > 0, \quad b \neq 0.$$

It appears that the latter satisfies the requirement (4.2.1) for the whole definition domain, i.e. for all $a > 0$, $b \neq 0$, such that $(ab+1) > 0$, whereas the measure $C^{(2)}_{n/m}(P/Q)$ given in (4.2.4) appears to be larger than $C_n^{(2)}(P)$ only for those a,b which guarantee its convexity in P defined by (4.1.34) (Van der Lubbe, 1981). The measure of conditional certainty given in (4.2.7) satisfies (4.2.1) for all $a > 0$. Some other properties common for all conditional certainty measures introduced in (4.2.4), (4.2.7), (4.2.8) are listed below (we omit the upper index in their notations to indicate that a given property holds for all of them).

1. Continuity

   $C_{n/m}(P/Q)$ is a continuous function of all q_j, p_{ij}, $i = 1,\ldots,n$; $j = 1,\ldots,m$. (4.2.9)

2. Symmetry

$C_{n/m}(P/Q)$ is a symmetric function with respect to q_j and p_{ij}, $j = 1,\ldots,m$; $i = 1,\ldots,n$. (cf. (1.1.20)). (4.2.10)

3. Expansibility (cf. 1.1.21).

$$C_{n+1/m}(P,0/Q) = C_{n/m}(P/Q). \qquad (4.2.11)$$

4. Monotony (cf. (4.1.30))

$C_{n/m}(P/Q)$ is a monotonically non-increasing function of p_{ij}, $i = 1,\ldots,n$; $j = 1,\ldots,m$. (4.2.12)

5. Minimality (cf. (4.1.27)) (4.2.13)

$$C_{n/m}(P/Q) \geq \frac{1}{n^a},$$

with equality iff $p_{ij} = \frac{1}{n}$ for all $i=1,\ldots,n$ and $j=1,\ldots,m$.

6. Maximality (cf.(4.1.26))

$$C_{n/m}(P/Q) \leq 1, \qquad (4.2.14)$$

with equality iff there is such an $i=k$ that $p_{kj}=1$ and $p_{ij}=0$ for all $j=1,\ldots,m$ and for all $i \neq k$.

Inequalities involving $C^{(2)}_{n/m}(P/Q)$ and $C^{(2a)}_{n/m}(P/Q)$ as well as limiting of properties of conditional certainty measures are given in (Van der Lubbe, 1981). Some other important properties will be discussed later.

The joint certainty measures can be derived from (4.1.20) and (4.1.22) by setting $r_{ij} = q_j p_{ij}$ $(i=1,\ldots,n;\ j=1,\ldots,m)$ instead of p_i.

Following this approach, we obtain

$$C^{(1)}_{nm}(P,Q) = \prod_{i=1}^{n} \prod_{j=1}^{m} (q_j p_{ij})^{aq_j p_{ij}}, \quad a > 0 \qquad (4.2.15)$$

and

$$C_{nm}^{(2)*}(P,Q) = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} (q_j p_{ij})^{ab+1} \right)^{\frac{1}{b}}, \quad a > 0, \ (ab+1) > 0, \ b \neq 0 \tag{4.2.16}$$

The joint certainty $C_{nm}^{(1)}(P,Q)$ related to the marginal certainty $C_{n}^{(1)}(P)$ given in (4.1.20) and to the conditional certainty $C_{n/m}^{(1)}(P,Q)$ given in (4.2.7) appears to be a symmetric function of P and Q,

$$C_{nm}^{(1)}(P,Q) = C_{nm}^{(1)}(Q,P).$$

It can be shown that the following desirable relations hold

$$C_{nm}^{(1)}(P,Q) = C_{m}^{(1)}(Q) \quad C_{n/m}^{(1)}(P/Q) \tag{4.2.17}$$

(strong multiplicativity),

$$C_{nm}^{(1)}(P,Q) \geq C_{n}^{(1)}(P) \quad C_{m}^{(1)}(Q) \tag{4.2.18}$$

(submultiplicativity)

$$C_{nm}^{(1)}(P,Q) = C_{n}^{(1)}(P) \quad C_{m}^{(1)}(Q), \tag{4.2.19}$$

iff P and Q are independent (weak multiplicativity).

Regarding the measure $C_{nm}^{(2)*}(P,Q)$ given in (4.2.16), it is also symmetric but does not possess the multiplicativity properties above, either in conjunction with $C_{n/m}^{(2)}(P/Q)$ as defined in (4.2.4), or with $C_{n/m}^{(2a)}(P/Q)$ defined by (4.2.8) (Van der Lubbe, 1981). Therefore, it cannot be considered as an appropriate measure for a joint certainty.

Van der Lubbe has introduced two measures corresponding to $C_{n/m}^{(2)}(P/Q)$ and $C_{n/m}^{(2a)}(P/Q)$ defined by

$$C_{nmm}^{(2)}(P,Q) = C_{m}^{(2)}(Q) \quad C_{n/m}^{(2)}(P/Q), \tag{4.2.19}$$

$$C_{nm}^{(2a)}(P,Q) = C_{m}^{(2)}(Q) \quad C_{n/m}^{(2a)}(P/Q). \tag{4.2.20}$$

It is obvious that the latter two measures have been designed so that the strong multiplicativity property should be automatically

guaranteed. Taking into consideration inequality (4.2.1) which is satisfied by both $C^{(2)}_{n/m}(P/Q)$ and $C^{(2a)}_{n/m}(P/Q)$, we can also see that the corresponding joint measures of certainty are sub-multiplicative, as defined in (4.2.18), and weak multiplicative, as defined in (4.2.19). But both of them appear to be non-symmetric and thus do not satisfy (4.2.17).

Certainty measures can serve as an appropriate basis for entropies. Since entropy is a measure of uncertainty, it should be a monotonically decreasing function of certainty. Three kinds of such functions have been suggested by Van der Lubbe (1981). Analogously to probability measures, all certainty measures discussed above take on values in $[0,1]$. As we have seen, the maximum value of certainty is achieved when one of the probabilities in a given distribution is equal to unit, while all other probabilities vanish. The minimum value appears to be dependent on a certain parameter of the certainty measure, as well as on the number of probabilities in the distribution. In the limiting case, however, as this number tends to infinity, all certainty measures tend to zero.

If we are willing to let the measure of uncertainty take on values in $[0,\infty]$ (as, e.g. the Shannon measure does), one of the non-linear monotonically decreasing functions has to be adopted. A conventional logarithmic function

$$H_n(P) = - \log C_n(P) \qquad (4.2.21)$$

or a hyperbolic function

$$H_n(P) = \frac{1}{C_n(P)} - 1 \qquad (4.2.22)$$

may be suitable in this case. If the entropy is supposed to take on values in $[0,1]$, we can have a linear function

$$H_n(P) = 1 - C_n(P). \qquad (4.2.23)$$

Regarding the additivity properties, the function given in (4.2.21) is the only one to yield weak additive entropies. Both linear and hyperbolic functions can only provide weak additivity of type β (see chapter 1).

5. INFORMATION MEASURES FOR CONTINUOUS DISTRIBUTIONS

Information measures considered in the preceeding chapters involve discrete probability distributions. However, all of them can be easily extended to continuous distributions as well. An appropriate generalization preserves the fundamental properties of the measure in question, though certain specific problems with respect to the continuity may arise. Information measures still exist, which have been developed specially for continuous probability distributions.

In the present chapter, we shall consider some of these information measures, especially those which seem to be most appropriate for estimation and identification purposes. We shall mainly concentrate on Kullback-Leibler's divergence and on Fisher's measure of information. Several generalizations of these two measures will be also discussed briefly later.

5.1 Kullback-Leibler divergence

The directed divergence discussed in section 2.1 has been originally introduced for continuous probability distributions (Kullback, 1951). We first consider its definition as given in (Kullback, 1951; 1959) and then proceed to the properties. Suppose X is a random variable, whose probability distribution f(x) is unknown. Suppose further that we have two hypotheses: H_1, implying that the true density function is $f_1(x)$ and H_2, implying that this density function is $f_2(x)$. Let $P(H_1)$ and $P(H_2)$ denote prior probabilities and $P(H_1/x)$, $P(H_2/x)$ denote posterior (after observation of one value X=x) probabilities of H_1 and H_2. It can be seen from the Bayes theorem that the following relation holds

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1/x)}{P(H_2/x)} - \log \frac{P(H_1)}{P(H_2)} \qquad (5.1.1)$$

The right hand side of (5.1.1) is a change of the log-likelihood ratio in favour of H_1 against H_2 for one observation X=x. The left hand side shows that this change is completely determined by the ratio of $f_1(x)$ and $f_2(x)$, i.e. by the degree of dissimilarity between H_1 and H_2 for a given x. The more H_1 and H_2 differ from each other, the more information for discrimination between them is provided by the given observation x. Taking expectations on both sides of equation (5.1.1) with respect to $f_1(x)$ results in

$$\int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx = \int f_1(x) \log \frac{P(H_1/x)}{P(H_2/x)} dx - \log \frac{P(H_1)}{P(H_2)} \qquad (5.1.2)$$

the right hand side of which is a mean (expected) increase of the log-likelihood ratio in favour of H_1 against H_2 for one observation, provided H_1 is true. It can also be interpreted as a mean (or expected) information for one observation in favour of H_1 against H_2, provided H_1 is true. Therefore, the left hand side of (5.1.2) is frequently termed information measure. Actually it is a measure of deviation of probability distribution $f_2(x)$ from the true probability distribution $f_1(x)$. In order to avoid any confusion with the Shannon measure of information, we shall refer to

$$J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = J\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx \qquad (5.1.3)$$

as the Kullback-Leibler directed divergence between $f_1(x)$ and $f_2(x)$. A symmetric divergence between $f_1(x)$ and $f_2(x)$ can be defined by

$$J(f_1,f_2) = J\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + J\begin{bmatrix} f_2 \\ f_1 \end{bmatrix} = \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} dx. \qquad (5.1.4)$$

In terms of hypotheses testing, it can be considered as the difference between the mean posterior log-likelihood ratio in favour of $H_1(H_2)$ against $H_2(H_1)$ after one observation, provided $H_1(H_2)$ is true, and the mean value of the said ratio, provided the hypothesis $H_2(H_1)$ is true.

The basic properties of the Kullback-Leibler directed divergence are identical to those of the Shannon directed divergence defined in (2.1.1). These are

1. Non-negativity ((cf. 2.1.5))

$$J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} > 0, \tag{5.1.5}$$

with equality iff $f_1(x) = f_2(x)$.

2. Strong additivity (cf.(2.1.10))

Let $f_1(x,y)$, $f_2(x,y)$ denote the joint probability densities of two random variables, X and Y. The the following equality holds

$$J\begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix} = J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} + \int f_1(x) \int f_1(y/x) \log \frac{f_1(y/x)}{f_2(y/x)} dy\, dx, \tag{5.1.6}$$

where $f_1(x) = \int f_1(x,y)\, dy$ and $f_2(x) = \int f_2(x,y)\, dy$ are the marginal probability densities of X and

$$f_1(y/x) = \frac{f_1(x,y)}{f_1(x)}, \quad f_2(y/x) = \frac{f_2(x,y)}{f_2(x)}$$

denote the conditional probability densities of Y given x.

3. Weak additivity (cf. (2.1.11))

$$J\begin{bmatrix} f_1(x) & f_1(y) \\ f_2(x) & f_2(y) \end{bmatrix} = J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} + J\begin{bmatrix} f_1(y) \\ f_2(y) \end{bmatrix}. \tag{5.1.7}$$

4. Inequalities

$$J\begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix} > J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}, \tag{5.1.8}$$

with equality iff

$$\int f_1(x) \int f_1(y/x) \log \frac{f_1(y/x)}{f_2(y/x)}\, dy\, dx = 0.$$

$$J\begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix} > \int f_1(x) \int f_1(y/x) \log \frac{f_1(y/x)}{f_2(y/x)}\, dy\, dx, \tag{5.1.9}$$

with equality iff $J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = 0$. Note that (5.1.8) and (5.1.9) also hold if we interchange x and y on the right hand side.

5. Invariance

Let $T(x)$: $X \to Y$ be a statistic, i.e. a measurable function whose domain and range are X and Y, respectively. Then the following relation holds

$$J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} \geq J\begin{bmatrix} f_1(t(x)) \\ f_2(t(x)) \end{bmatrix}, \qquad (5.1.10)$$

where $f_1(t(x))$ and $f_2(t(x))$ are the probability densities of $T(x)$ corresponding to those of X, with equality iff

$$\frac{f_1(x)}{f_2(x)} = \frac{f_1(t(x))}{f_2(t(x))}.$$

With respect to the symmetric divergence defined by (5.1.4) we have the following properties.

1. Non-negativity

$$J[f_1(x), f_2(x)] \geq 0, \qquad (5.1.11)$$

with equality iff $f_1(x) = f_2(x)$.

2. Symmetry

$$J[f_1(x), f_2(x)] = J[f_2(x), f_1(x)]. \qquad (5.1.12)$$

3. Strong additivity

$$J[f_1(x,y), f_2(x,y)] = J[f_1(x), f_2(x)] + \qquad (5.1.13)$$
$$+ \int [f_1(x,y) - f_2(x,y)] \log \frac{f_1(y/x)}{f_2(y/x)} \, dy \, dx$$

4. Weak additivity

$$J[f_1(x)\, f_1(y), f_2(x)\, f_2(y)] = J[f_1(x), f_2(y)] +$$
$$+ J[f_1(y), f_2(y)]. \qquad (5.1.14)$$

5. Inequalities

$$J[f_1(x,y),\ f_2(x,y)] \geq J[f_1(x),\ f_2(x)], \qquad (5.1.15)$$

with equality iff $\int [f_1(x,y) - f_2(x,y)] \log \frac{f_1(y/x)}{f_2(y/x)}\, dy\, dx = 0.$

$$J[f_1(x,y),\ f_2(x,y)] \geq \int [f_1(x,y) - f_2(x,y)] \log \frac{f_1(y/x)}{f_2(y/x)}\, dy\, dx, \qquad (5.1.16)$$

with equality iff $J[f_1(x),\ f_2(x)] = 0.$

Again the same relations hold if we interchange x and y on the right hand side.

6. Invariance

$$J[f_1(x),\ f_2(x)] \geq J[f_1(t(x)),\ f_2(t(x))], \qquad (5.1.17)$$

with the same notations and the same condition for equality as in (5.1.10). A statistic, for which (5.1.10) and (5.1.17) become equalities, is said to be sufficient. It is sufficient in the sense that it preserves all information in X, which is relevant for discrimination between $f_1(x)$ and $f_2(x)$.

We can also introduce joint and conditional divergences. For two random variables X and Y, the directed joint divergence is given by

$$J\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}; X,Y = J\begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix} = \int f_1(x,y) \log \frac{f_1(x,y)}{f_2(x,y)}\, dx\, dy \qquad (5.1.18)$$

and the symmetric joint divergence can be defined by

$$J[f_1,f_2;X,Y] = J[f_1(x,y),\ f_2(x,y)] = \int [f_1(x,y) - f_2(x,y)] \log \frac{f_1(x,y)}{f_2(x,y)}\, dx\, dy. \tag{5.1.19}$$

The directed conditional divergence

$$J\left|\begin{matrix} f_1 \\ f_2 \end{matrix};\ Y/X\right| = \int f_1(x)\ J\left|\begin{matrix} f_1(y/x) \\ f_2(y/x) \end{matrix}\right| dx = \int f_1(x,y) \log \frac{f_1(y/x)}{f_2(y/x)}\, dx\, dy \tag{5.1.20}$$

and the symmetric condition divergence

$$J[f_1,\ f_2;\ Y/X] = \int [f_1(x,y) - f_2(x,y)] \log \frac{f_1(y/x)}{f_2(y/x)}\, dx\, dy \tag{5.1.21}$$

are suggested by (5.1.6) and (5.1.13), respectively.

The properties 2 to 4 of the directed divergence can be expressed in terms of the joint and conditional measures (5.1.18) and (5.1.20) as follows.

2a. Strong additivity (cf. (5.1.6))

$$J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X,Y\right] = J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X\right] + J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ Y/X\right].$$

3a. Weak additivty (cf. (5.1.7))

$$J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X,Y\right] = J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X\right] + J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ Y\right]$$

iff X and Y are independent.

4a. Inequalities

$$J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X,Y\right] \geq J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X\right], \text{ with equality iff } J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ Y/X\right] = 0$$

(cf. (5.1.8)).

$$J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X,Y\right] \geq J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ Y/X\right], \text{ with equality iff } J\left[\begin{matrix} f_1 \\ f_2 \end{matrix};\ X\right] = 0$$

(cf. (5.1.9)).

In the same way we could reformulate the basic properties of the

symmetric divergence in terms of the joint and conditional measures given in (5.1.19) and (5.1.21). The property given in (5.1.7) (weak additivity) implies that the mean directed divergence, expected from N independent observations $X^1,\ldots,X^N$, is N times greater than that expected from one observation provided, of course, $f(x^1) = \ldots = f(x^N) = f(x)$. We thus have

$$J\left[\begin{matrix} f_1 \\ f_2 \end{matrix} ; X^1,\ldots,X^N\right] = N\, J\left[\begin{matrix} f_1 \\ f_2 \end{matrix} ; X\right]. \qquad (5.1.22)$$

Suppose we have two hypotheses: H_1, implying $f_1(x)$, and H_2, implying $f_2(x)$. Let H_2 denote the null hypothesis. Then the probability α_N of a wrong accepting H_1 (the type I error) on the basis of N independent observations on X and the probability β_N of a wrong accepting H_2 (the type II error) can be related to the Kullback-Leibler divergence by

$$N\, J\left[\begin{matrix} f_1 \\ f_2 \end{matrix} ; X\right] \geq \beta_N \log \frac{\beta_N}{1-\alpha_N} + (1-\beta_N) \log \frac{1-\beta_N}{\alpha_N} \qquad (5.1.23)$$

and

$$N\, J\left[\begin{matrix} f_2 \\ f_1 \end{matrix} ; X\right] \geq \alpha_N \log \frac{\alpha_N}{1-\beta_N} + (1-\alpha_N) \log \frac{1-\alpha_N}{\beta_N} \qquad (5.1.24)$$

(Kullback, 1959). These inequalities give the lower bound on α_N (β_N) for the given $f_1(x)$, $f_2(x)$, N and $\beta_N(\alpha_N)$.

It should be emphasised that (5.1.10) and (5.1.17) impose an upper bound upon the performance of an estimator T(x) in terms of divergence for two probability densities $f_1(x)$, $f_2(x)$. The relation of the Kullback-Leibler directed divergence to other information measures will be discussed in the next chapter.

5.2 Fisher information

Perhaps the very first probabilistic information measure was introduced

by Fisher (1925). The Fisher measure of information is intrinsically based on parametric probability distributions of continuous variables and is specially designed for statistical estimation. In the present section we shall consider the definition of Fisher's measure and its basic properties.

Suppose the probability distribution of some continuous random variable X is given by a density function $f(x;\theta)$ dependent on a scalar parameter θ. For a given value of θ, $f(x;\theta)$ becomes a deterministic function of x only. If θ is unknown, it can be estimated through observations on X on the basis of the given relationship $f(x;\theta)$. Each observation on X provides a certain information about the value of θ, and this information is entirely determined by the function $f(x;\theta)$. The stronger the dependence of $f(x;\theta)$ on θ, the larger the information concerning the value of θ provided by the given observation $X = x$.

In the limiting case, when $f(x;\theta)$ is such that there exists a one-to-one correspondence between x and θ, one single observation on X gives the true value of θ and thus provides exhaustive information. Another limiting case arises when f does not depend on θ. Certainly, in such a case, observations on X cannot yield any information about θ.

Suppose the density function $f(x;\theta)$ satisfies the following regularity conditions:

1. $\dfrac{\partial f(x;\theta)}{\partial\theta}$ exists for all values of θ and almost all x; (5.2.1)

2. $$\frac{\partial}{\partial\theta}\int_G f(x;\theta)\,dx = \int_G \frac{\partial f(x;\theta)}{\partial\theta}\,dx, \qquad (5.2.2)$$

   for all values of θ and for all G belonging to the definition domain of $f(x;\theta)$;

3. $$\int \left[\frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)}\right]^2 f(x;\theta)\,dx < \infty. \qquad (5.2.3)$$

The mean Fisher information about a certain value θ of Θ expected from one observation on x is given by

$$F(\theta;\ X) = \int \left[\frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)}\right]^2 f(x;\theta)\,dx =$$

$$= \int \left[\frac{\partial \ln f(x;\theta)}{\partial \theta} \right]^2 f(x;\theta)\, dx. \qquad (5.2.4)$$

It can be seen from (5.2.4) that the Fisher information about θ depends on the true value of θ. For a given density function of θ denoted by $f(\theta)$, the mean Fisher information about the true value of θ expected from one observation on X can be evaluated by

$$F(\Theta;X) = \int F(\theta;X)\, f(\theta)\, d\theta. \qquad (5.2.5)$$

The regularity conditions (5.2.1) to (5.2.3) are satisfied e.g. for a so-called exponential (Koopman-Darmois) family of density functions given by

$$f(x;\theta) = C(\Theta) \exp \left[\sum_j S_j(\Theta)\, T_j(x) \right] g(x), \qquad (5.2.6)$$

where $C(\Theta)$ and $S_j(\Theta)$ are certain functions of Θ alone (independent of x) and $T_j(x)$, $g(x)$ are functions of x alone (independent of Θ). This is not the case, however, when the range of the density function $f(x;\theta)$ depends on θ and the tails of $f(x;\theta)$ do not touch the x-axis (see e.g. (Papaioannou, 1970)).

Suppose $f(x,y;\theta)$ is the joint probability density of two random variables X and Y, satisfying the regularity conditions (5.1.1) to (5.1.3). The mean Fisher information about the true value θ of Θ expected from one observation on both X and Y is given by

$$F(\theta;X,Y) = \iint \left[\frac{\partial \ln f(x,y;\theta)}{\partial \theta} \right]^2 f(x,y;\theta)\, dx\, dy, \qquad (5.2.7)$$

called the joint Fisher information. Averaging $F(\theta,X,Y)$ with respect to the probability distribution $f(\theta)$ of Θ results in

$$F(\Theta;X,Y) = \int F(\theta;X,Y)\, f(\theta)\, d\theta, \qquad (5.2.8)$$

which is a measure of the Fisher information about Θ expected from one observation on both X and Y.

Let us consider now the conditional Fisher measure of information. The

mean Fisher information about the true value θ of Θ expected from one observation on Y given x can be measured by

$$F(\theta;Y/x) = \int \left[\frac{\partial \ln f(y/x;\theta)}{\partial\theta}\right]^2 f(y/x;\theta)\, dy, \qquad (5.2.9)$$

where

$$f(y/x;\theta) = \frac{f(x,y;\theta)}{\int f(x,y;\theta)\, dy}$$ is the conditional probability

density of Y given x. Taking expectation on both sides of (5.2.9) with respect to the probability distribution of X, $f(x;\theta) = \int f(x,y;\theta)\, dy$, gives rise to

$$F(\theta;Y/X) = \iint \left[\frac{\partial \ln f(y/x;\theta)}{\partial\theta}\right]^2 f(x,y;\theta)\, dx\, dy, \qquad (5.2.10)$$

which is a <u>conditional measure of Fisher's information</u>. It is a measure of the mean Fisher information about the true value θ of Θ provided by one observation on Y given an arbitrary value x of X. Again, averaging with respect to the probability distribution $f(\theta)$ of Θ gives the mean Fisher information about Θ provided by one observation on Y given arbitrary value x of X,

$$F(\Theta;Y/X) = \int F(\theta;Y/X)\, f(\theta)\, d\theta. \qquad (5.2.11)$$

Note that the Fisher information measures (5.2.5), (5.2.8) and (5.2.11) do not depend on the parameter Θ.

We can now state the properties of Fisher's information measure defined in (5.2.4), (5.2.7) and (5.2.10), which obviously hold for all values of θ of Θ.

1. Non-negativity

$$F(\theta;X) \geq 0, \qquad (5.2.12)$$

with equality iff $f(x;\theta)$ is independent of Θ.

2. Symmetry

$$F(\theta;X,Y) = F(\theta;Y,X). \qquad (5.2.13)$$

3. Strong additivity

$$F(\theta;X,Y) = F(\theta;X) + F(\theta;Y/X). \qquad (5.2.14)$$

4. Sub-additivity

$$F(\theta;X,Y) \leq F(\theta;X) + F(\theta;Y). \qquad (5.2.15)$$

5. Weak additivity

For independent X and Y it holds

$$F(\theta;X,Y) = F(\theta;X) + F(\theta;Y). \qquad (5.2.16)$$

6. Inequalities

$$F(\theta;X,Y) \geq F(\theta;X), \qquad (5.2.17)$$

with equality iff $F(\theta;Y/X) = 0$.

$$F(\theta;X,Y) \geq F(\theta;Y/X), \qquad (5.2.18)$$

with equality iff $F(\theta;X) = 0$.

$$F(\theta;Y/X) \geq F(\theta;Y), \qquad (5.2.19)$$

with equality iff X and Y are independent.

7. $$F(\theta;Y/X) = 0, \qquad (5.2.20)$$

iff X is sufficient for θ.

8. Let T be a measurable transformation of X with a density function $f(t;\theta)$. Then it holds

$$F(\theta;T) \leq F(\theta;X), \qquad (5.2.21)$$

with equality iff T is sufficient for θ.

9. Convexity

$F(\theta;X)$ is a convex function of $f(x;\theta)$. (5.2.22)

10. Cramér-Rao inequality (Rao, 1945)

Let $\tau(\theta)$ be a continuously differentiable function of θ. If $T = T(x)$ is an unbiased estimator of $\tau(\theta)$, then it holds (for all θ)

$$E[(T - \tau(\theta))^2] \geq \frac{[\frac{d}{d\theta}\tau(\theta)]^2}{F(\theta;X)}, \qquad (5.2.23)$$

where E denotes the mathematical expectation with respect to $f(x;\theta)$. If T is a biased estimator of θ with the bias $b(\theta)$, then we can write

$\tau(\theta) = \theta + b(\theta)$, and the latter inequality transforms to

$$E[(T - \theta)^2] \geq \frac{(1+ \frac{d}{d\theta} b(\theta))^2}{F(\theta;X)} \quad \text{for all } \theta. \qquad (5.2.24)$$

For an unbiased estimator $\tilde{e}$ of θ, (5.2.24) reduces to

$$E[(\tilde{\theta} - \theta)^2] \geq \frac{1}{F(\theta;X)} \quad \text{for all } \theta, \qquad (5.2.25)$$

which gives a lower bound on the variance of an unbiased estimator.

Suppose, for instance, $f(x;\theta)$ is a Gaussian function with unknown mean θ and variance σ^2. The Fisher information about θ expected from on observation on X can be evaluated by (5.2.4), which yields $F(\theta;X) = \frac{1}{\sigma^2}$. Let us take $\tilde{\theta} = x$ as an estimator for θ. The variance of $\tilde{\theta}$, $E[(\tilde{\theta} - \theta)^2] = E[(x - \theta)^2] = \sigma^2$, i.e. is equal to $\frac{1}{F(\theta;X)}$, which is the right hand side of (5.2.25). It follows thus that $\tilde{\theta} = x$ is the best possible estimator of the mean θ, based on on observation x. For N independent observations $x_1,\ldots,x_N$ the Fisher (joint) information becomes (by (5.2.16))

$$F(\theta;X_1,\ldots,X_N) = N\, F(\theta;X) = \frac{N}{\sigma_2}.$$

If we take the sample mean $\frac{1}{N}\sum_i x_i$ as an estimator $\tilde{\theta}$, we find $E[(\tilde{\theta} - \theta)^2] = \frac{\sigma^2}{N}$ and thus again (5.2.25) becomes equality, implying that in this case the sample mean is an estimator which possesses the least possible variance.

A generalization of the Cramér-Rao inequality due to Boekee (1977) is given by

$$E[|T - \tau(\theta)|^s] \geq \frac{\left|\frac{d}{d\theta}\tau(\theta)\right|^s}{F_s(\theta;X)}, \quad s > 1, \text{ for all } \theta, \qquad (5.2.26)$$

where

$$F_s(\theta;X) = \left[\int \left|\frac{\partial \ln f(x;\theta)}{\partial\theta}\right|^{\frac{s}{s-1}} f(x;\theta)\,dx\right]^{s-1}, \quad s > 1 \tag{5.2.27}$$

is a generalized Fisher information measure. For s = 2, (5.2.26) and (5.2.27) reduce to (5.2.23) and (5.2.4), respectively. Inequality (5.2.26) gives an upper bound on the performance of T in terms of the s-th absolute moment of its error.

Suppose Θ is a random parameter with a probability density $f(\theta)$. Making use of (5.2.5), we can obtain inequality (Gart, 1959)

$$E_\theta[E[(\tilde{\theta} - \theta)^2]] \geq \frac{1}{F(\Theta;X)}, \tag{5.2.28}$$

where E_θ denotes expectation with respect to $f(\theta)$. This inequality is an extension of the Cramér-Rao inequality given in (5.2.25).

A similar extension can also be written for (5.2.26). Several modifications of the Cramér-Rao inequality stating sharper bounds on the performance of an estimator are available (Barankin, 1949; Bhattacharyya, 1946; Chapman 1951).

As we have already mentioned, the Fisher information depends, in general, on the value of the parameter Θ. This dependence is determined by the function $f(x;\theta)$. In this respect, two kinds of parameters, the so-called location parameters and scale parameters (Kagan, 1973; Boekee, 1978) show peculiarity. Let Θ be a location parameter of X, i.e. such that there exists an $Y = X - \Theta$ with probability density $f(y)$ satisfying

$$f(x;\Theta) = f(y). \tag{5.2.29}$$

Then it can be shown that the Fisher information does not depend on the value of Θ, so that $F(\theta;X) = F(0;X) = F(\theta;X)\big|_{\theta=0}$.
If Θ is a scale parameter of X, i.e. such that there exists an $Y = \frac{X}{\Theta}$ with probability density $f(y)$ satisfying

$$f(x;\Theta) = \frac{1}{\Theta} f(y), \tag{5.2.30}$$

then $F(\theta;X) = \frac{1}{\theta^2}\ F(1;X) = F(\theta;X)\Big|_{\theta=1} \frac{1}{\theta^2}$ (Kagan, 1973; Boekee, 1978).

An extension of the Fisher information measure given in (5.2.4) to a vector valued parameter $\Theta = (\theta_1,\ldots,\theta_n)$ is given by a symmetric matrix

$$F(\Theta;X) = \left\|F(\theta_i,\theta_j;X)\right\|,\ i=1,\ldots,n;\quad j=1,\ldots,n \tag{5.2.31}$$

with

$$F(\theta_i,\theta_j;X) = \int \frac{\partial \ln f(x;\Theta)}{\partial\theta_i}\ \frac{\partial \ln f(x;\Theta)}{\partial\theta_j}\ f(x;\Theta)\,dx, \tag{5.2.32}$$

$i=1,\ldots,n;\quad j=1,\ldots,n$. Note that the diagonal elements of the Fisher information matrix, $F(\theta_i,\theta_i;X)$, $i=1,\ldots,n$, are equal to the Fisher information measures $F(\theta_i;X)$, $i=1,\ldots,n$ for respective components of Θ, while the meaning of other elements is somewhat vague.
The Cramér-Rao inequality suggests, however, a reasonable interpretation: $F(\theta_i,\theta_j;X)$ with $i \neq j$ gives a lower bound on the covariance for the respective estimators $\tilde{\theta}_i$ and $\tilde{\theta}_j$.

A simple modification of (5.2.31) has been introduced by Papaioannou (1970). Neglecting the off-diagonal elements $F(\theta_i,\theta_j;X)$, $i \neq j$, we can define the Fisher information about the value θ of a vector parameter Θ by

$$F^*(\Theta;X) = \sum_{i=1}^{n} F(\theta_i;X), \tag{5.2.33}$$

i.e. as the trace of the Fisher information matrix (5.2.31). The joint modified Fisher information can be given by

$$F^*(\Theta;X,Y) = \sum_{i=1}^{n} F(\theta_i;X,Y) \tag{5.2.34}$$

and the conditional modified Fisher information can be defined as a sum

$$F^*(\Theta;Y/X) = \sum_{i=1}^{n} F(\theta_i;Y/X). \tag{5.2.35}$$

The mean Fisher information about a vector parameter Θ can be expressed as a matrix

$$F(\theta;X) = \left\| F(\theta_i,\theta_j;X) \right\|, \quad i=1,\ldots,n; \quad j=1,\ldots,n, \qquad (5.2.36)$$

where

$$F(\theta_i,\theta_j;X) = E_{\theta_i,\,\theta_j}[F(\theta_i,\theta_j;X)]$$

$$= \iint F(\theta_i,\theta_j;X)\ f(\theta_i,\theta_j)\ d\theta_i\, d\theta_j, \qquad (5.2.37)$$

$i=1,\ldots,n$; $j=1,\ldots,n$, $f(\theta_i,\theta_j)$ being the joint probability density of the respective components of θ.

The modified mean Fisher information about θ is given by the trace of $F(\theta;X)$,

$$F^*(\theta;X) = \sum_{i=1}^{n} F(\theta_i,\theta_j;\ X) \qquad (5.2.38)$$

Some further extensions of the Fisher information measure will be discussed in the next section. The relation of Fisher's information to other information measures is considered in chapter 6.

5.3 Generalizations of Fisher's information measure

Several generalizations of the Fisher measure of information are known, both imposing upon $f(x;\theta)$ less stringent regularity conditions and improving the Cramér-Rao inequality. In this section we shall consider contributions to the former of the two problems mentioned above. One of the first generalizations of this kind is due to Kagan (1963). Kagan's measure of information is based on the χ^2-divergence of one probability distribution, say $f_2(x)$, from another distribution, $f_1(x)$, which is defined by

$$\chi^2\left[\frac{f_1(x)}{f_2(x)}\right] = \int \left[\frac{f_1(x) - f_2(x)}{f_1(x)}\right]^2 f_1(x)\, dx. \qquad (5.3.1)$$

Let $f(x;\theta)$ denote the probability density of X with a parameter θ and $f(x;\theta+\Delta\theta)$ be the same density function of X with $\theta+\Delta\theta$. Setting $f_1(x) = f(x;\theta)$ and $f_2(x) = f(x;\theta+\Delta\theta)$ in (5.3.1) leads to the following

definition of the generalized Fisher-Kagan information measure:

$$FK(\theta;X) = \lim_{\Delta\theta\to 0} \inf \frac{1}{(\Delta\theta)^2} \chi^2 \begin{bmatrix} f(x;\theta) \\ f(x;\theta+\Delta\theta) \end{bmatrix}. \qquad (5.3.2)$$

it is obvious that $FK(\theta;X)$ exists even when $f(x;\theta)$ is not differentiable with respect to θ and thus appears to be free from regularity assumptions given in (5.2.1) to (5.2.3). If the density function $f(x;\theta)$ happens to satisfy the regularity conditions, (5.3.2) reduces to the usual Fisher information as defined in (5.2.4). It can easily be proved that all basic properties of the Fisher information measure listed in (5.2.12) to (5.2.25) above remain valid for the generalized Fisher-Kagan measure as well. Inequality (5.2.23) attains the form of

$$E[(T - \tau(\theta))^2] \geq \frac{\left[\frac{d}{d\theta}\tau(\theta)\right]^2}{FK(\theta;X)}, \qquad (5.3.3)$$

where T is an unbiased estimator of $\tau(\theta)$.

For a multivariate parameter $\Theta = (\Theta_1,\ldots,\Theta_n)$ the following relation

$$\left|\left|\tilde{\theta}\right|\right| - \left|\left|FK(\theta;X)\right|\right|^{-1} \geq 0 \qquad (5.3.4)$$

gives an extension of the Cramér-Rao inequality (5.2.25), where $\left|\left|\tilde{\theta}\right|\right|$ is the covariance matrix of the estimator $\tilde{\theta} = (\tilde{\theta}_1,\ldots,\tilde{\theta}_n)$ and

$$\left|\left|FK(\theta;X)\right|\right| = \left|\left|FK(\theta_i,\theta_j;X)\right|\right| \qquad (5.3.5)$$

with

$$FK(\theta_i,\theta_j;X) = \lim_{|\Delta\theta|\to 0} \inf \frac{1}{\Delta\theta_i \Delta\theta_j} \int \frac{f(x;\theta)-f(x;\theta+\Delta\theta_i)}{f(x;\theta)} \cdot$$

$$\cdot \frac{f(x;\theta)-f(x;\theta+\Delta\theta_j)}{f(x;\theta)} f(x;\theta)\, dx. \qquad (5.3.6)$$

A further generalization of the Fisher information measure, which can also be considered as an extension of the Fisher-Kagan information, has been suggested by Vajda (1973). The generalization consists of a more general measure of divergence termed χ-divergence of order α being used, which is given by

$$\chi^{\alpha}\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \int \left| \frac{f_1(x) - f_2(x)}{f_1(x)} \right|^{\alpha} f_1(x)\, dx, \quad \alpha > 1. \tag{5.3.7}$$

The generalized Fisher-Vajda information measure is defined by

$${}_{\alpha}FV(\theta;X) = \lim_{\Delta\theta\to 0} \inf \frac{1}{|\Delta\theta|^{\alpha}} \chi^{\alpha}\begin{bmatrix} f(x;\theta) \\ f(x;\theta+\Delta\theta) \end{bmatrix}, \quad \alpha > 1. \tag{5.3.8}$$

Setting $\alpha = 2$, we obtain the Fisher-Kagan information (5.3.2). The measure ${}_{\alpha}FV(\theta;X)$ appears to be non-negative and additive. It also possesses an invariance property so that if $T(X)$ is a measurable function of X, then it holds

$${}_{\alpha}FV(\theta;X) \geq {}_{\alpha}FV(\theta;T(X)), \text{ for all } \theta, \tag{5.3.9}$$

with equality iff T is sufficient for θ. An extension of the Cramer-Rao inequality (5.2.23) in terms of ${}_{\alpha}FV(\theta;X)$ is given by (Vajda, 1973)

$$\left\{ E\left[\left| T-\tau(\theta) \right|^{\beta} \right] \right\}^{\frac{1}{\beta}} \geq \frac{\frac{d}{d\theta}\tau(\theta)}{\left[{}_{\alpha}FV(\theta;X) \right]^{\alpha-1}} \quad \alpha > 1 \tag{5.3.10}$$

$\frac{1}{\alpha} + \frac{1}{\beta} = 1$. This relation states a lower bound on the βth absolute mean error of an unbiased estimator T for $\tau(\theta)$. With $\alpha = \beta = 2$ it reduces to

$$E\left[(T-\tau(\theta))^2 \right] \geq \frac{\left[\frac{d}{d\theta}\tau(\theta) \right]^2}{{}_2FV(\theta;X)}, \tag{5.3.11}$$

which is a generalization of (5.2.23) in the sense that it is valid irrespective of the regularity conditions (5.2.1) to (5.2.3). Note that the latter expression is identical to (5.3.3), since ${}_2FV(\theta;X) = FK(\theta;X)$.

A further generalization, based on Csiszár's ϕ-divergence (Csiszár, 1967), has been suggested by Aggarwal (1974). Let $f_1(x)$ and $f_2(x)$ denote probability densities of X. The directed ϕ-divergence of the density function $f_2(x)$ from the density function $f_1(x)$ is defined by

$$J^*_{\phi}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \int f_2(x) \quad \phi\left[\frac{f_1(x)}{f_2(x)}\right] dx, \tag{5.3.12}$$

where $\phi(u)$ is a convex function of u satisfying certain regularity conditions (see (Csiszár, 1967)). Note that for $f_1(x) = f_2(x)$, this measure is equal to $\phi(1)$. Suppose ϕ is a non-negative function. Then it can easily be seen that the "improved" directed ϕ-divergence given by

$$J_{\phi}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = J^*_{\phi}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} - \phi(1) =$$

$$= \int f_2(x) \quad \phi\left[\frac{f_1(x)}{f_2(x)}\right] dx \quad - \phi(1). \tag{5.3.13}$$

is a non-negative measure satisfying

$$J_{\phi}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \geq 0. \tag{5.3.14}$$

Setting $\phi(u) = u\log u$, we obtain from (5.3.13) the Kullback-Leibler directed divergence of $f_2(x)$ from $f_1(x)$ as defined in (5.1.3), and $\phi(u) = u(1 - \frac{1}{u})^2$ yields the χ^2-divergence given in (5.3.1).

The more general χ-divergence of order α can be derived as a special case of (5.3.13) with $\phi(u) = u(1 - \frac{1}{u})^{\alpha}$, which appears to be identicial to (5.3.7).

Substituting $f(x;\theta)$ for $f_1(x)$ and $f(x;\theta+\Delta\theta)$ for $f_2(x)$ in (5.3.13), we btain the directed ϕ-divergence of $f(x;\theta+\Delta\theta)$ from $f(x;\theta)$,

$$J_{\phi}\begin{bmatrix} f(x;\theta) \\ f(x;\theta+\Delta\theta) \end{bmatrix} = \int f(x;\theta+\Delta\theta) \quad \phi\left[\frac{f(x;\theta)}{f(x;\theta+\Delta\theta)}\right] dx - \phi(1). \tag{5.3.15}$$

following Aggarwal (1974), we can define the generalized Fisher information based on ϕ-divergence as

$$F_{\phi}(\theta;X) = \lim_{\Delta\theta\to 0} \inf \frac{1}{(\Delta\theta)^2} \quad J_{\phi}\begin{bmatrix} f(x;\theta) \\ f(x;\theta+\Delta\theta) \end{bmatrix}. \tag{5.3.16}$$

This measure can be considered as a further generalization of the Fisher-Kagan information given in (5.3.2), which follows from (5.3.16) for $\phi(u) = u(1 - \frac{1}{u})^2$. Let T be a measurable function of X. Then the following relation holds

$$F_\phi(\theta;T(X)) \geq F_\phi(\theta;X), \tag{5.3.17}$$

with equality iff T is sufficient for θ. This relation is clearly a generalization of (5.2.21). The relation of $F_\phi(\theta;X)$ with the Fisher information $F(\theta;X)$ has been discussed in (Aggarwal, 1974). It should be mentioned that $F_\phi(\theta;X)$ cannot be considered as a generalization of ${}_\alpha FV(\theta;X)$ as defined in (5.3.8), though both measures become identical for $\phi(u) = u(1 - \frac{1}{u})^\alpha$ with $\alpha = 2$.

An extension of the generalized Fisher information measure defined in (5.2.27), based on the ϕ-divergence, has been introduced by Boekee (1977).

A different approach has been suggested by Papaioannou (1970). As we have seen from the considerations above, the probabilistic measures of divergence do not require any regularity conditions except (for some of them, e.g. for Kullback-Leibler's divergence) that the probability measures involved should be absolutely continuous with respect to one another. Consider two density functions, $f(x;\theta)$ and $f(\theta;\ \theta+\Delta\theta)$, and suppose $\Delta\theta$ is a known (finite) change of θ, dependent on θ, such that

$$\Delta\theta = q(\theta) - \theta,$$

where q denotes a one-to-one mapping of Θ onto itself, satisfying

$$q(\theta) \neq \theta. \tag{5.3.18}$$

The directed divergence of $f(x;\theta+\Delta\theta) = f(x;q(\theta))$ from $f(x;\theta)$ appears to be a function of both θ and q. If we fix q, the divergence becomes a function of θ alone. It can be regarded thus as a measure of information about θ expected from one observation on X. Suppose q is

such that $f(x;\theta) = 0$ whenever $f(x;q(\theta)) = 0$, which implies that the corresponding probability measures are absolutely continuous with respect to one another. In this case, the directed Kullback-Leibler measure of divergence appears to be applicable. Setting $f(x;\theta)$ for $f_1(x)$ and $f(x;q(\theta))$ for $f_2(x)$ in (5.1.3) yields

$$J\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix}; X = J\begin{bmatrix}f(x;\theta) \\ f(x;q(\theta))\end{bmatrix} = \int f(x;\theta) \log \frac{f(x;\theta)}{f(x;q(\theta))}\, dx, \tag{5.3.19}$$

called the modified Kullback-Leibler functional information measures (based on q).

This measure can also be considered as a generalization of the Fisher information, since it is free from regularity assumptions, which appear in the definition of the latter, and at the same time it is essentially a measure of information about a parameter θ.

For two random variables X and Y with a given joint probability density $f(x,y;\theta)$ we can introduce

$$J\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix}; X,Y = J\begin{bmatrix}f(x,y;\theta) \\ f(x,y;q(\theta))\end{bmatrix} = \int f(x,y;\theta) \log \frac{f(x,y;\theta)}{f(x,y;q(\theta))}\, dxdy \tag{5.3.20}$$

called the joint modified Kullback-Leibler functional information measure based on q. For a given q, it can be considered as a measure of information about θ expected from one observation on both X and Y. The conditional modified Kullback-Leibler functional information measure based on q is given by

$$J\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix}; Y/X = J\begin{bmatrix}f(x;\theta) \\ f(x;q(\theta))\end{bmatrix}; Y/X = \int f(x,y;\theta) \log \frac{f(y/x;\theta)}{f(y/x;q(\theta))}\, dx\, dy, \tag{5.3.21}$$

which follows from (5.1.20) with $f_1(y/x) = f(y/x;\theta)$ and $f_2(y/x) = f(y/x;q(\theta))$. Being a special case of the Kullback-Leibler directed divergence, the measure given in (5.3.19) to (5.3.21) possesses all the properties listed in (5.1.5) through (5.1.10). The latter attains the

following form

$$J\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix}; X\Big] \geq J\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix}; T\Big], \quad \text{for all } \theta, \tag{5.3.22}$$

with equality iff t is sufficient for θ.

The modified Kullback-leibler information measure based on q is related to the Fisher information measure by

$$\lim_{q(\theta)\to\theta} \frac{J\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix}; X\Big]}{(q(\theta) - \theta)^2} \approx \frac{1}{2} F(\theta;X) \tag{5.3.23}$$

(see (Papaioannou, 1970)). An extension to multivariate as well as to discrete parameter θ can be easily obtained. Note that in a multivariate case (5.3.19) remains a scalar, which may be considered as another advantage over the Fisher information measure.

The modified Kullback-leibler functional information measure is not applicable, however, if the range of the probability density $f(x;\theta)$ depends on the parameter θ. In such a case, there exists no transformation $q(\theta)$ satisfying (5.3.18) and such that $f(x;\theta) = 0$ whenever $f(x;q(\theta)) = 0$. This difficulty may be overcome by making of use different measure of dissimilarity between $f(x;\theta)$ and $f(x,q(\theta))$, which would be insensitive to the latter condition. An appropriate solution, as suggested by Papaioannou (1970), can be obtained on the basis of the Bhattacharyya divergence (Bhattacharyya, 1943)

$$\rho[f_1(x), f_2(x)] = \int\sqrt{f_1(x)\, f_2(x)}\, dx \tag{5.3.24}$$

or by its generalization

$$\rho^*[f_1(x), f_2(x)] = \int\sqrt[s]{f_1(x)}\ \sqrt[t]{f_2(x)}\, dx, \quad s > 0, \ t > 0, \tag{5.3.25}$$

$\frac{1}{s} + \frac{1}{t} = 1$ is called affinity. For $s = t = 2$ affinity becomes symmetric with respect to $f_1(x)$ and $f_2(x)$. Affinity ranges over $[0,1]$ with $\rho^* = 0$ for mutually singular density functions $f_1(x)$, $f_2(x)$ and $\rho^* = 1$ for identical distributions, $f_1(x) = f_2(x)$. It is also known to possess the invariance property given by

$$\rho^*[f_1(x),\ f_2(x)] = \rho^*[f_1(T(x)),\ f_2(T(x))], \qquad (5.3.26)$$

where $T(x)$ is a measurable transformation of X, which is sufficient for X, and $f_1(T(x))$, $f_2(T(x))$ are the probability densities induced by this transformation.

The Bhattacharyya functional information measure based on q can be defined by

$$B\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix};X\Big] = -\ln \int [f(x;\theta)]^{\frac{1}{s}}\ [f(x,q(\theta))]^{\frac{1}{t}}\ dx =$$

$$= -\ln \int \left[\frac{f(x;\theta)}{f(x;q(\theta))}\right]^{\frac{1}{s}}\ f(x;q(\theta))\ dx, \qquad (5.3.27)$$

where $s > 0$, $t > 0$, $\frac{1}{s} + \frac{1}{t} = 1$ and $q(\theta)$ is a one-to-one mapping of Θ onto itself satisfying the only condition given in (5.3.18).

The joint and conditional Bhattacharyya functional information measures based on q are given by

$$B\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix};X,Y\Big] = -\ln \iint \left[\frac{f(x,y;\theta)}{f(x,y;q(\theta))}\right]^{\frac{1}{s}} f(x,y;q(\theta))\ dx\ dy \qquad (5.3.28)$$

and

$$B\begin{bmatrix}\theta \\ q(\theta)\end{bmatrix};X/Y\Big] = -\ln \iint \left[\frac{f(x/y;\theta)}{f(x/y;q(\theta))}\right]^{\frac{1}{s}} f(x,y;q(\theta))\ dx\ dy \qquad (5.3.29)$$

respectively.

The following properties of the Bhattacharyya functional information measure can be easily obtained from its definition given in (5.3.27) to (5.3.29) (see also (Papaioannou, 1970)).

1. Non-negativity

$$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right] \geq 0 \text{ for all } \theta, \qquad (5.3.30)$$

with equality iff $f(x;\theta)$ does not depend on θ.

2. Weak additivity

$$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X,Y\right] = B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right] + B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; Y\right], \qquad (5.3.31)$$

iff X and Y are independent and q is such that $f(x;\theta) = 0$ for all x, for which $f(x;q(\theta)) = 0$ and $f(y;\theta) = 0$ for all y, for which $f(y;q(\theta)) = 0$.

3. Non-(strong) additivity

$$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X,Y\right] \not= B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right] + B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; Y/X\right]. \qquad (5.3.32)$$

4. Invariance

$$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; T(X)\right] \leq B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right] \qquad (5.3.33)$$

with equality iff T is sufficient for θ.

5. $$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; Y/X\right] \geq 0, \qquad (5.3.34)$$

with equality iff X is sufficient for θ.

6. $$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; Y/X\right] \geq B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; Y\right], \qquad (5.3.35)$$

with equality iff X and Y are independent.

Suppose $q(\theta) > \theta$ for all θ. Then the Bhattacharyya functional information measure can be related to the Fisher information measure by

$$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right] \simeq -\ln\left[1 - \frac{(q(\theta)-\theta)^2}{2st} F(\theta;X)\right] \qquad (5.3.36)$$

The relation to the modified Kullback-Leibler functional information measure is given by the following inequality

$$B\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right] \leq \frac{J\left[\begin{matrix}\theta \\ q(\theta)\end{matrix} ; X\right]}{t} \qquad (5.3.37)$$

(Papaioannou, 1970). The relations (5.3.30) through (5.3.37) hold, of course, for all θ.

It should also be mentioned that both the Kullback-Leibler functional measure and the Battacharyya functional measure fail to be additive with respect to the components of a multivariate parameter so that information about the value of a vector θ does not equal the sum of information evaluated for each component of θ.

6. INFORMATION-THEORETIC APPROACH TO IDENTIFICATION

Identification is a cognitive process whose principal task is the acquisition of knowledge about an object. The information-theoretic approach thus seems to be a good basis for discussing methods/techniques for deriving objective criteria and for presenting a unified picture or identification in general. The present chapter is an attempt to give an account of useful and effective applications of information theory in identification and to indicate some new problems, which can be solved within an information-theoretic framework.

The first section deals with the measures of information in identification, with especial emphasis on their mutual relations in the context of estimation problems. Restriction has been made to the Shannon, Fisher and Kullback-Leibler information measures, which seem to be the most appropriate ones from the point of view of identification. We shall dwell upon the Kerridge measure of inaccuracy, closely related to the above-mentioned information models.

In the subsequent sections, an information-theoretic approach to identification will be developed and numerous applications will be discussed. Special attention is paid to the problems, the solution of which in the information-theoretic framework seems to be the most effective or even the only possible one, such as:

- determination of the most informative input test signals;
- choice of the probability distribution and, consequently, of the identification method most adequate to the a priori knowledge;
- determination of structure (order) of the model.

Numerous contributions to the solution of traditional problems such as prediction, filtering, smoothing made by information approach are also discussed. It appears that all optimal identification methods and techniques are derived through an appropriate information approach.

6.1 Information in identification

Most of the phenomena dealt with in identification are stochastic and continuous. Therefore, the best suitable information models seem to be those which are based on the prior knowledge represented by continuous probability distributions. In this respect the following distinction appears to be useful. Prior probabilities may be given as deterministic functions of random variables alone, or as functions of those and of certain unknown parameters which, in turn, can be deterministic or random quantities. In what follows we shall refer to parametric or non-parametric probability distributions, dependent on whether unknown parameters are involved or not. Information models based on probability density functions will be termed parametric or non-parametric, repsectively. The Shannon and Kullback-Leibler measures, for instance, are non-parametric, whereas Fisher's information is essentially parametric. As we have seen in section 5.3, parametric modification of non-parametric models can be easily derived. An inverse transition is hardly possible: parametric models lose their value when losing unknown parameters.

Information in an experiment is usually regarded as a reduction of uncertainty due to the experiment. From this point of view, the Shannon theory suggests perhaps one of the most appropriate models for identification. The Shannon measure of uncertainty for discrete

probability distributions has been already discussed in the first chapter.

Let us now consider, in brief, its continuous analogue, which will be used later as a basis for an information model of identification.

Suppose X is a real valued random variable with probability density f(x). Then uncertainty about the true instantaneous value of X can be measured by

$$H(X) = -\int f(x) \log f(x)\, dx, \qquad (6.1.1)$$

called the (Shannon) entropy of X. Uncertainty about the true instantaneous values of two random variables X, Y with a bivariate density function f(x,y) is given by the joint entropy

$$H(X,Y) = -\iint f(x,y) \log f(x,y)\, dx\, dy. \qquad (6.1.2)$$

The mean uncertainty about the value of Y given the value of X is determined by the conditional entropy

$$H(Y/X) = -\iint f(x,y) \log f(y/x)\, dx\, dy, \qquad (6.1.3)$$

where f(y/x) is the conditional probability density of Y given X = x. The basic properties of the Shannon entropy defined in (6.1.1) to (6.1.3) are identical to the respective properties of its discrete analogue presented in section 1.1. The most useful of them is the strong additivity, which in this case attains the form

$$H(X,Y) = H(X) + H(Y/X) = H(Y) + H(X/Y). \qquad (6.1.4)$$

For stochastically independent variables X and Y we have H(Y/X) = H(Y) and (6.1.4) reduces to

$$H(X,Y) = H(X) + H(Y), \qquad (6.1.5)$$

termed weak additivity.

Suppose now that X is an unobservable random variable and Y is an observable random variable related to X by the known joint probability density f(x,y). The mean information about the true instantaneous

value of X expected from one single observation on Y is measured by the mean (expected) reduction in uncertainty about X due to observation on Y,

$$I(X;Y) = H(X) - H(X/Y), \tag{6.1.6}$$

where H(X/Y) is the conditional entropy of X given Y, defined, analogously to (6.1.3), by

$$H(X/Y) = -\int f(x,y) \log f(x/y)\, dx\, dy. \tag{6.1.7}$$

Substituting (6.1.1) and (6.1.7) in (6.1.6) results in the Shannon measure of information

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy. \tag{6.1.8}$$

This measure is symmetric in X,Y and non-negative, i.e.

$$I(X;Y) = I(Y;X) \geq 0. \tag{6.1.9}$$

It is also additive so that for two independent observed variables Y_1, Y_2 we have

$$I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2) \tag{6.1.10}$$

The Shannon information as defined in (6.1.8) is a non-parametric measure, though it becomes parametric if the probability densities involved are given as functions of an unknown parameter.

Let us now related the Shannon information and the Kullback-Leibler directed divergence given in (5.1.3). Setting $f_1(x) = f(x,y)$ and $f_2(x) = f(x)f(y)$ in (5.1.3) yields

$$J\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = J\begin{bmatrix} f(x,y) \\ f(x)f(y) \end{bmatrix} = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy, \tag{6.1.11}$$

the right hand side of which is exactly the right hand side in (6.1.8).

It follows thus that in this particular case the Kullback-Leibler directed divergence reduces to the Shannon information. Note that equality $f_2(x) = f(x)f(y)$ implies that X and Y are stochastically independent. In terms of hypotheses testing, the Kullback-Leibler divergence (6.1.11) can be expressed (on account of (5.1.2)) by

$$J\begin{bmatrix} f(x,y) \\ f(x)f(y) \end{bmatrix} = \int f(x,y) \log \frac{P(H_1/x,y)}{P(H_2/x,y)} \, dx \, dy - \log \frac{P(H_1)}{P(H_2)} , \tag{6.1.12}$$

where H_1 implies that the joint probability density of X and Y is $f_1(x,y) = f(x,y)$ and H_2 implies another density function $f_2(x,y) = f(x)f(y)$ with $f(x) = \int f_1(x,y)\, dy = \int f(x,y)\, dy$ and $f(y) = \int f_1(x,y)\, dx = \int f(x,y)\, dx$. It appears thus that the marginal probability densities of X and Y under both H_1 and H_2 are the same. The only difference that exists is that H_2 considers X and Y as independent variables, whereas H_1 ascribes to them a certain dependence given by $f(x,y)$.

It follows from (6.1.8) and (6.1.11) that the mean Shannon information about the value of X expected from one observation on Y, given the stochastic dependence between X and Y, can be regarded as the mean directed Kullback-Leibler divergence of the joint distribution $f_2(x,y)$ from the true distribution $f_1(x,y) = f(x,y)$, provided $f_2(x,y)$ is a particular case of $f_1(x,y)$ with independent X and Y, so that $f_2(x,y) = f(x)\, f(y) = [\int f(x,y)\, dy][\int f(x,y)\, dx]$.

Another interpretation of the relation between Shannon's information and Kullback-Leibler's divergence arises by equating (6.1.8) and (6.1.12):

$$I(X,Y) = \int f(x,y) \log \frac{P(H_1/x,y)}{P(H_2/x,y)} \, dx \, dy - \log \frac{P(H_1)}{P(H_2)} . \tag{6.1.13}$$

It is apparent from (6.1.13) that the mean Shannon information about the value of X expected from one observation on Y given $f(x,y)$ is equal to the mean (expected) increase of the log-likelihood ratio in favour of hypothesis H_1, implying the true density function $f(x,y)$, against another hypothesis H_2, implying a different density function $f_2(x,y) = f(x)f(y)$ (by preserving the true marginal distributions of X and Y but proclaiming them independent), which could be obtained through one observation on both X and Y, provided, of course, that X is also observable.

Another information measure which seems to be very appropriate for estimation purposes is the Kerridge inaccuracy discussed in section 3.1. The continuous analogue of (3.1.1) is given by

$$\bar{A}\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = -\int f_1(x) \log f_2(x) \, dx, \qquad (6.1.14)$$

called inaccuracy of the probability density $f_2(x)$ with respect to the probability density $f_1(x)$. The following equality can be easily obtained from (6.1.1), (5.1.3) and (6.1.14)

$$\bar{A}\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = -\int f_1(x) \log f_1(x) \, dx + \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \, dx =$$

$$= H(f_1(x)) + J\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} \qquad (6.1.15)$$

(cf. (3.1.2)), which relates inaccuracy to both Shannon's entropy and Kullback-Leibler's directed divergence. Inaccuracy (6.1.14) appears to be a non-negative and strongly additive measure. Suppose $f_1(x)$ represents the true probability density of a random variable X due to its intrinsic stochastic nature and $f_2(x)$ is an approximation of $f_1(x)$, based on some inaccurate knowledge of X. Then inaccuracy can be regarded as a measure of total uncertainty about the true instantaneous value of X due to both intrinsic vagueness of X (given by $f_1(x)$) and to inaccuracy of our knowledge (given by $f_2(x)$) about this vagueness. Note that the minimal value of inaccuracy is bounded by the entropy $H(f_1(x))$ and is achieved iff $f_2(x) = f_1(x)$.

We can also relate the Kerridge inaccuracy to the Shannon entropy and information directly, without involving the directed divergence. Setting $f_1(x) = f(x,y)$ and $f_2(x) = f(x)f(y)$ in (6.1.14), one gets

$$\bar{A}\begin{bmatrix} f(x,y) \\ f(x)f(y) \end{bmatrix} = -\int f(x,y) \log f(x) \, dx \, dy - \int f(x,y) \log f(y) \, dx \, dy$$

$$= -\int f(x) \log f(x) \, dx - \int f(y) \log f(y) \, dy =$$

$$= H(X) + H(Y), \qquad (6.1.16)$$

which coincides with the right hand side in (6.1.5). This implies that the Kerridge inaccuracy of a hypothesis, stating that the the joint

probability density of X and Y is $f_2(x,y) = f(x)f(y) = [\int f(x,y)\,dy]\,[\int f(x,y)\,dx]$ (i.e. assuming the true marginal probability densities of X and Y but claiming them independent), while the true joint density is f(x,y), appears to be identical with the Shannon joint entropy of X and Y evaluated under the conditions stipulated by this hypothesis (i.e. for independent X and Y with the true marginal probability distributions).

Taking into consideration (6.1.16), (6.1.6) and (6.1.4), we can obtain another relation

$$I(X;Y) = A\left[\frac{f(x,y)}{f(x)f(y)}\right] - H(X,Y), \qquad (6.1.17)$$

from which it follows that the Shannon information about X expected from one observation on Y, given the true joint probability density f(x,y), is equal to the difference between the inaccuracy of a hypothesis, implying that the joint probability density is $f_2(x,y) = f(x)f(y)$ (i.e. assuming X and Y to be independent variables with the true marginal probability densities f(x), f(y)) and the Shannon joint entropy of X and Y, corresponding to their true joint probability density f(x,y).

Incidentally, relations (6.1.16) and (6.1.17) suggest another interpretation of the Shannon information. As mentioned above, Shannon's information obtained from a certain message is defined as a reduction in uncertainty (in terms of the Shannon entropy), which is due to this message. The second term on the right hand side of (6.1.17) can be regarded as a measure of uncertainty about the true values of both X and Y after receipt of a message about their true joint probability density f(x,y), whereas the first term, given by the right hand side of (6.1.16), is a measure of uncertainty about the true values of X and Y, which corresponds to an incorrect prior knowledge implying independence of X and Y.

We can state now that the Shannon information about the true value of X

expected from one observation on Y, given their (true) joint prob ability density f(x,y), is equal to the Shannon information gain from a message about the true stochastic relationship between X and Y, corresponding to a prior knowledge, which implies the true marginal probability distributions of X and Y but incorrectly assumes them to be independent. In a case where X and Y are, in fact, independent, such a message does not yield any new knowledge, and the right hand side of (6.1.17) becomes 0. Consequently, $I(X/Y) = 0$, which means that certainly observations on Y cannot give any information about X, since Y is independent of X.

The relationship between Fisher's information and Kullback-Leibler's divergence has been discussed in (Kullback, 1959). Let $f(x;\theta)$ be a density function of a certain random variable with a multivariate parameter $\theta = (\theta_1,\ldots,\theta_n)$. Consider another density function $f(x;\theta+\Delta\theta)$ differing from $f(x;\theta)$ by a slight change of its parameter, $\Delta\theta = (\Delta\theta_1,\ldots,\Delta\theta_n)$. Setting $f_1(x) = f(x;\theta)$ and $f_2(x) = f(x;\theta+\Delta\theta)$ in the right hand side of (5.1.3) and making use of the Taylor series expansion results in the following relation between Kullback-Leibler's directed divergence of $f(x;\theta+\Delta\theta)$ from $f(x;\theta)$ and Fisher's information about θ expected from one observation on X

$$J\begin{bmatrix} f(x;\theta) \\ f(x;\theta+\Delta\theta) \end{bmatrix} \simeq \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} F(\theta_i,\theta_j;X)\, \Delta\theta_i \Delta\theta_j, \qquad (6.1.18)$$

with $F(\theta_i,\theta_j;X)$ as defined in (5.2.32).

For a scalar parameter θ, the above relation reduces to

$$J\begin{bmatrix} f(x;\theta) \\ f(x;\theta+\Delta\theta) \end{bmatrix} \simeq \frac{1}{2} (\Delta\theta)^2 F(\theta;X), \qquad (6.1.19)$$

implying that a small change $\Delta\theta$ of a scalar parameter θ results in a Kullback-Leibler directed divergence of the density function from its initial form $f(x;\theta)$, which can be approximated by the product of the Fisher information about θ expected from one observation on X and a factor $\frac{1}{2}(\Delta\theta)^2$.

Now we are in a postition to relate the Fisher and Shannon measures of information. Suppose $f(x;y;\theta)$ is a joint probability density of two

random variables X, Y with a scalar parameter θ. The mean Fisher information about a certain value θ of Θ expected from one observation on both X and Y given $f(x,y;\theta)$ can be expressed by

$$F(\theta;X,Y) = \iint f(x,y;\theta) \left[\frac{\partial \ln f(x,y;\theta)}{\partial\theta}\right]^2 dx\, dy \qquad (6.1.20)$$

(cf. (5.2.7)).

Consider

$$\iint f(x,y;\theta) \log \frac{f(x,y;\theta)}{f(x,y;\theta+\Delta\theta)} dx\, dy, \qquad (6.1.21)$$

where $\Delta\theta$ is such a change of the parameter Θ that

$$f(x,y;\theta+\Delta\theta) = \left[\int f(x,y;\theta)\, dy\right]\left[\int f(x,y;\theta)\, dx\right] = f(x;\theta)f(y;\theta),$$

i.e. it preserves the marginal densities of both X and Y but makes them independent. Then (6.1.21) can be regarded as the mean Kullback-Leibler directed divergence of $f(x,y;\theta+\Delta\theta)$ from $f(x,y;\theta)$ as defined in (5.1.3) and, at the same time, as the mean Shannon information (6.1.8) about the true value of X expected from one observation on Y given $f(x,y;\theta)$. Now on account of (6.1.19) we have

$$I(X;Y) = \frac{1}{2}(\Delta\theta)^2 F(\theta;X,Y), \qquad (6.1.22)$$

which relates the Shannon and Fisher measures of information.

It follows thus that the mean Shannon information about the true value of X expected from one observation on Y given $f(x,y;\theta)$ is approximately equal to the mean Fisher information about the value of the parameter θ expected from one observation on both X and Y (provided, of course, X is also observable), given $f(xy;\theta)$, multiplied by a factor $\frac{1}{2}(\Delta\theta)^2$, where $\Delta\theta$ is such a change of Θ that X and Y become independent with the same marginal probability densities.

If Θ is a vector valued parameter $(\Theta_1,\ldots,\Theta_n)$, then by (6.1.18)

$$I(X;Y) \approx \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} F(\theta_i,\theta_j;X,Y)\ \Delta\theta_i \Delta\theta_j \qquad (6.1.23)$$

can be obtained in the same manner, where $\Delta\theta = (\Delta\theta_1,\ldots,\Delta\theta_n)$ is such a change of Θ that X and Y become independent with the same marginal densities, and

$$F(\theta_i,\theta_j;X,Y) = \int \frac{\partial \ln f(x,y;\theta)}{\partial \theta_i} \frac{\partial \ln f(x,y;\theta)}{\partial \theta_j} f(x,y;\theta)\, dx\, dy$$

(cf.(5.2.32)).

6.2 Information-theoretic estimation principle

Information measures appear to be suitable performance criteria in estimation and identification. Surely, an estimate $\tilde{a}$ of a model parameter or an estimate $\tilde{x}$ of a certain unobservable variable x should provide as much information about the true value a or x as possible. For parametric models given in a form of a density function $f(x;\theta)$, the best estimate $\tilde{\theta}$ of an unknown parameter θ, given observations on X, can be obtained by use of the Fisher information theory. If the optimal estimator appears to be computationally intractable, we can choose a feasible estimator whose performance (in terms of the Fisher information), related to the best possible performance following from the Cramér-Rao inequality, shows information loss due to its non-optimality.

For non-parametric models, when the relationship between the observed and the estimated variables can be expressed in the form of a certain equation rather than by a parametric probability distribution, the direct application of the Fisher information theory fails. In such cases the Shannon measure of information involving non-parametric distributions appears to provide a more appropriate performance criterion. Let us consider estimation of an unobservable variable X through observations on another variable, say Y. The mean information about the true value of X expected from one observation on Y given the joint probability density $f(x,y)$ is determined by (6.1.8). Let $\tilde{X} = \tilde{X}(Y)$ denote and estimate of X based on observations on Y, let $f(x,\tilde{x})$ be the joint probability density of X and $\tilde{X}$ and let $f(x)$,

$f(\tilde{x})$ be the joint probability densities of X and $\tilde{X}$, respectively. Then the Shannon information about X provided by $\tilde{X}$ is given by

$$I(X;\tilde{X}) = \iint f(x,\tilde{x}) \log \frac{f(x,\tilde{x})}{f(x)f(\tilde{x})} \, dx \, d\tilde{x}, \qquad (6.2.1)$$

which follows directly from (6.1.8).

It can be shown that the following relation holds

$$I(X;\tilde{X}) \leq I(X;Y), \qquad (6.2.2)$$

with equality iff $\tilde{X}$ is sufficient for X, i.e. preserves all information in Y about X. Regarding $I(X;\tilde{X})$ as a performance criterion for $\tilde{X}$, we arrive at the basic information-theoretic estimation principle: choose that estimate $\tilde{X} = \tilde{X}(Y)$, which yields the largest information $I(X;\tilde{X})$ defined in (6.2.1). Thus optimizing $\tilde{X}$ reduces to maximizing $I(X;\tilde{X})$.

Information $I(X;\tilde{X})$ can be expressed through the entropy analogously to (6.1.6),

$$I(X;\tilde{X}) = H(X) - H(X/\tilde{X}), \qquad (6.2.3)$$

where H(X) and $H(X/\tilde{X})$ denote the marginal entropy of X and the conditional entropy of X given $\tilde{X}$, respectively. It follows from (6.2.3) that maximizing $I(X;\tilde{X})$ is equivalent to minimizing $H(X/\tilde{X})$. The information-theoretic estimation principle implies then a choice of such an estimate $\tilde{X}$ so that the condition entropy

$$H(X/\tilde{X}) = -\int f(x,\tilde{x}) \log f(x/\tilde{x}) \, dx \, d\tilde{x} \qquad (6.2.4)$$

attains its minimum. it can also be shown that the following relation holds

$$H(X/\tilde{X}) \geq H(X/Y), \qquad (6.2.5)$$

with equality iff $\tilde{X}$ is sufficient for X.

In order to relate the information principle to the well-established estimation principles, let us consider the estimate error

$$\hat{X} = \tilde{X} - X. \qquad (6.2.6)$$

It is apparent from (6.2.6) that the probability density of $\hat{X}$ is equal to the conditional probability density of X given $\tilde{X}$,

$$f(\hat{x}) = f(x-\tilde{x}) = f(x/\tilde{x}) \qquad (6.2.7)$$

and thus,

$$H(\hat{X}) = H(X/\tilde{X}). \qquad (6.2.8)$$

Now it is clear that minimizing $H(X/\tilde{X})$ is equivalent to minimizing $H(\hat{X})$ and the said information-theoretic estimation principle can be reformulated as follows: choose that estimate $\tilde{X}$, for which the error entropy is minimal. The following relation

$$H(\hat{X}) \geq H(X/Y), \qquad (6.2.9)$$

with equality iff $\tilde{X}$ is sufficient for X, can be obtained from (6.2.5) and (6.2.8).

The latter formulation of the information-theoretic principle seems to be illuminating for its relation to other estimation principles and methods. The error entropy may be written as

$$H(\hat{X}) = -\int f(\hat{x}) \log f(\hat{x})\, d\hat{x} = -E[\log f(\hat{x})] \qquad (6.2.10)$$

which is a kind of a moment of the error $\hat{X}$. The least squares method, for instance, is based on minimizing the second central moment of the error (variance). It is known that if $f(\hat{x})$ is Gaussian, then $H(\hat{X})$ is entirely determined by the variance of $\hat{X}$ alone, and thus minimizing $H(\hat{X})$ implies minimizing the variance. In this case the information-theoretic principle reduces to the well-known least squares method.

On the other hand, the least squares method is known to be a particular kind of the maximum likelihood principle under the same condition, i.e. by a normal distribution. Consequently, the information principle is equivalent to the maximum likelihoodprinciple, provided the observations are normally distributed.

Yet another formulation of the information-theoretic estimation principle is possible. It was first introduced by Weidemann (1969; 1970) in the context of control systems analysis. A measure of information in $\hat{X}$ about Y appears to be a good performance criterion for $\tilde{X}$. The following relation

$$I(Y;\tilde{X}) = H(\hat{X}) + H(Y) - H(\hat{X},Y), \qquad (6.2.11)$$

which can be derived from (6.1.4) and (6.1.6) after setting Y for X and $\hat{X}$ for Y, together with

$$H(\hat{X},Y) = H(\tilde{X}-X,Y) = H(\tilde{X}(Y)-X,Y) = H(X,Y), \qquad (6.2.12)$$

which is a direct consequence of (6.1.2) and (6.2.6), lead to

$$\begin{aligned} I(Y;\hat{X}) &= H(\hat{X}) + H(Y) - H(X,Y) = \\ &= H(\hat{X}) - H(X/Y), \end{aligned} \qquad (6.2.13)$$

relating information in X about Y and the error entropy. It follows hence that minimizing $H(\hat{X})$ is equivalent to minimizing $I(Y;\hat{X})$, since H(X/Y) is a constant for a given probability density f(x,y).

In terms of $I(Y;\hat{X})$, the information-theoretic estimation principle implies that <u>the estimate $\tilde{X}$ should minimize information about the observed variable Y provided by the estimate error $\hat{X}$.</u> The lower bound on $I(Y;\hat{X})$ is given by

$$I(Y;\hat{X}) \geq 0, \qquad (6.2.14)$$

with equality iff $\tilde{X}$ is sufficient for X, which is a direct consequence of (6.2.9) and (6.2.13).

It is obvious that $I(X;\tilde{X})$, $H(X/\tilde{X})$, $H(\hat{X})$ and $I(Y;\hat{X})$ are equivalent performance criteria for $\tilde{X}$: optimizing either of them results in the same estimate $\tilde{X}$. It is only a matter of convenience as to which criterion out of (6.2.1), (6.2.4), (6.2.10) and (6.2.13) to use in a given problem.

In a case where the probability distributions involved in information

criteria are unknown and only the respective variances are available, we can assume, by the maximum entropy principle, (see next section), the Gaussian densities yielding the maximal entropies $H(X/\tilde{X})$ and $H(\hat{X})$. Implementation of the information-theoretic estimation principle implies minimization of $H(X/\tilde{X})$ or $H(\hat{X})$, and this procedure is termed the minimax entropy principle. Note that by making use of the said minimax entropy principle Rissanen succeeds in estimating the structure of a statistical model (see section 6.4).

Several applications of the information-theoretic principle to identification and estimation, including recursive filtering, prediction and smoothing are known. Let e.g. $\tilde{X}(Y,K)$ denote a given form of a recursive estimator with an unknown parameter K. Maximizing $I(X;\tilde{X}(Y,K))$ or minimizing $H(\hat{X}(K))$ or $I(Y;\hat{X}(K))$ with respect to K results in an optimal estimate of K. Kalata and Priemer (1979) have shown that the maximum information recursive estimator coincides with the Kalman filter in the case of a linear Gaussian model. When the probability distribution is unknown, the use of the minimax entropy principle results in an estimator, which is also identical to a Kalman filter. This is no suprise, since the minimax entropy principle is equivalent to the least squares method used in the Kalman filtering. In (Kalata, 1979), the optimal linear recursive predictor and smoother have been derived by an information-theoretic principle. A computationally feasible minimax error entropy recursive filter for nonlinear systems (Kalata, 1974) compares favourably (especially for small numbers of observations) to an optimal mean square error filter, regarding the average absolute error $|\hat{X}|$. Filtering problems for linear discrete and continuous Gaussian processes have been considered by an information-theoretic approach also in (Tomita, 1976; 1976a).

The idea of identification by an information approach can be explained as follows. Let us consider a model $Y = Y(X,A)$, which relates an observable variable Y to an unobservable variable X, A being an unknown (also unobservable) parameter. The problem consists of estimating A through observations on Y. The use of an information-theoretic principle implies minimizing the estimate error entropy $H(\hat{A}) = H(\tilde{A}-A)$, which results in an optimal estimate $\tilde{A}$. For a given form

of estimator $\tilde{A} = \tilde{A}(Y,K)$ minimizing $H(\hat{A}(K)) = H(\tilde{A}(Y,K) - A)$ with respect to K leads to an optimal estimate for K. A recursive minimax error entropy estimator $\tilde{A}(Y,K)$ (Kalata, 1978) shows much resemblance to an adaptive identification algorithm based on the Kalman filtering theory.

The information-theoretic approach also seems to suggest the best way to choose an optimum input test signal. Certainly, such an optimal signal should provide maximum information about the estimated process. Suppose we have a model $Y = Y(X,A)$, where X,Y and A are an input variable, an output variable and an unknown parameter, respectively. The problem is to determine (under certain constraints) the best form of the input signal X for identification of A by observations on Y. Maximizing information $I(A;Y(X))$ about A, provided by Y, with respect to X under the given constraints results in an optimal input signal $\tilde{X}$. For a given shape of the input signal $X = X(K)$, maximizing $I(A;Y(K))$ with respect to K, gives an optimum value of the unknown parameter K.
Arimoto and Kimura (1971) implemented an information-theoretic approach to design the optimal input test signals for identification of impulse response function as well as of transfer function of a linear system with a white Gaussian noise under a constraint imposed on the signals amplitude ($X \leq C$). In both cases a pseudo-random binary sequence X with an amplitude C appears to be an optimal information-theoretic solution.

6.3 Prior probability distributions in identification

There is no need to prove the importance of a choice of the probability distribution related to an estimated process. The whole theory of identification is essentially based on a probabilistic approach.
It is also well recognized that, unfortunately, very often the prior knowledge available does not specify the probability distribution

needed for the choice of an appropriate identification model. Sometimes the shape of a probability density is suggested by the physical nature of the estimated process. It frequently happens, however, that even such "pre-knowledge" is lacking. In such cases the decisions taken are usually based on subjective (sometimes explicit) assumptions.
Information theory provides a good means to overcome this arbitrariness. It seems reasonable to choose that specific probability distribution which corresponds to the least certainty and, consequently, to the least information expected from observations, under a given prior knowledge. In discussing this problem we shall make use of the distinction introduced in the first section of the present chapter. Two kinds of statistical models - parametric and non-parametric - imply two different forms of relationship between the observed and estimated variables, corresponding to two different information models.

Let us consider first the choice of a non-parametric probability distribution. Suppose we are going to estimate an observable variable X through observations on another variable Y related to X. Suppose further that the true relationship between X and Y is unknown. We may have, however, some (vague) information about X or Y or even about their mutual relationship, which can be used in estimating X. For instance, X may be known to be a non-negative variable, contaminated by some noise, the probability distribution of which is unknown. It might seem that in this case the best estimate of X is the observation on Y itself. For negative values of Y, however, such a decision rule appears to be in conflict with the prior knowledge, implying that X is non-negative.

We can improve the estimate by taking into consideration the probability density of X, which certainly should satisfy the following conditions

$$f(x) = 0, \text{ for } x < 0.$$
$$\int f(x)dx = 1. \qquad (6.3.1)$$

The next step is to choose a density function f(x), which would not

assume any other knowledge of X and thus correspond to the least certainty (or to the maximal uncertainty) about X under the given prior knowledge. Since we are willing to use a non-parametric probability density f(x), the Shannon entropy appears to be an appropriate measure of uncertainty. These considerations lead to a so-called maximum entropy principle advocated by Jaynes (1968), implying the following decision rule: take that distribution f(x), which, under certain constraints (a priori knowledge), yields to the maximum entropy of X. Any other distribution chosen corresponds to some additional knowledge, which, as a matter of fact, does not exist, or neglects some prior knowledge available.

Suppose the prior knowledge on X to be given by the mean values of certain functions of X, say $q_i(x)$, i=1,...,k. Taking equations

$$\int q_i(x)\, f(x)\, dx = \overline{q_i(x)}, \quad i=1,\ldots,k \tag{6.3.2}$$

and certainly (6.3.1) as constraints and making use of the method of Lagrange multipliers, we can find that specific probability density f(x), which yields a maximal Shannon entropy defined in (6.1.1). The solution is given by (Kopilovich, 1964; Boekee, 1976)

$$f(x) = \exp(\lambda_o - 1 + \sum_{i=1}^{k} \lambda_i q_i(x)), \tag{6.3.3}$$

where λ_o, λ_i (i=1,...,k) are the constants, which can be found on account of constraints (6.3.1) and (6.3.2). If, for example, the prior knowledge is represented by a known variance σ^2 of X, then we have

$$\overline{q_1(x)} = \sigma^2$$

and

$$q_1(x) = (x - m)^2,$$

where m is the mean of X. Equation (6.3.3) gives in this case a Gaussian function f(x) with σ^2 as variance.

A different form of constraints arises when only the mean value of a sum

$$q(x) = \sum_{i=1}^{k} t_i \, q_i(x) \qquad (6.3.4)$$

is known. Maximizing the Shannon entropy (6.1.1) under constraints given in (6.3.1) and

$$\overline{q(x)} = \sum_{i=1}^{k} t_i \overline{q_i(x)} = \int q(x) \, f(x) \, dx \qquad (6.3.5)$$

results in (Boekee, 1976)

$$f(x) = \exp(\lambda_o - 1 + \lambda_1 \sum_{i=1}^{k} t_i \, q_i(x)), \qquad (6.3.6)$$

where $t_i (i=1,\ldots,k)$ are the constants given in (6.3.4) and λ_o, λ_1 are the Lagrange multipliers, which can be found by (6.3.1) and (6.3.5).

If the prior information on f(x) cannot be expressed in the form of (6.3.2) or (6.3.4), (6.3.5), different optimization procedures may be needed. As to our knowledge, no solutions for an optimal f(x) under such (more general) constraints are available so far.

An inverse problem arises in situations when, for certain reasons, we are inclined to assume a particular probability density f(x). Then we may ask what additional information on X is needed in order that this specific function should be justified. In other words, which constraints should be imposed upon X in order to make f(x) the most probable density function for X. The Gaussian distribution, for example, appears to be optimal iff the variance of X is known. A general solution for f(x) given in (6.3.3) or (6.3.6) can be expressed by (6.3.2) (Noonan, 1976) or by (6.3.5) (Rozenberg, 1966), respectively. The difference between these two solutions exists in the number of constraints.
If, for instance, f(x) is supposed to be a Rayleigh distribution

$$f(x) = a^2 \, x \exp\left(-\frac{a^2 x^2}{2}\right), \quad x \geq 0,$$

then making use of (6.3.2) results in two constraints given by

$$\overline{q_1(x)} = \overline{x^2} = \int x^2 \, f(x) \, dx$$

and

$$\overline{q_2(x)} = \overline{\ln x} = \int \ln x \, f(x) \, dx,$$

whereas (6.3.5) leads to one single constraint

$$\overline{q(x)} = \overline{\frac{x^2}{2}} - \overline{\ln x} = \int \left(\frac{x^2}{2} - \ln x\right) f(x) \, dx$$

(Noonan, 1976).

Let us consider now the choice of a parametric probability distribution. Suppose that the estimated variable can be considered as parameter θ of a probability density $f(x;\theta)$ of some observed variable X. The estimated parameter θ can be a scalar or vector, random or non-random variable. Both X and Θ may be continuous or discrete variables. We also assume that no functional relationship between X and Θ is known (otherwise we could use the non-parametric model discussed above). Suppose now that the probability density $f(x;\Theta)$ is unknown. We may still have some information about X and/or about Θ or, perhaps, about the influence of Θ upon the probability distribution of X. We may know, e.g., that Θ is a location parameter, satisfying (5.2.29), or a scale parameter, as defined in (5.2.30). It seems reasonable to accept that density function $f(x;\Theta)$, which, under constraints given by this prior knowledge, corresponds to the least certainty about Θ after one observation on X. Any other distribution chosen would neglect some knowledge available or imply some additional prior knowledge which, in fact, does not exist. An appropriate measure of certainty about Θ on the evidence of one observation on X is given by the Fisher information defined in (5.2.4), (5.2.5).

Now we are in a position to formulate the decision rule for the choice of an optimal prior parametric probability distribution, which we call a minimum information principle: take that probability density $f(x;\Theta)$ which, under certain constraints, determined by the prior knowledge on X and Θ, yields the minimal Fisher information about Θ expected from one observation on X.

This principle seems to be consistent with the maximum entropy

principle discussed above, from which it differs in two aspects: first, it applies to parametric probabilistic models and, secondly, it is based on the Fisher information measure, whereas the latter makes use of the Shannon entropy. The probability density $f(x;\Theta)$ chosen by using the minimum information principle can be considered as a basis for the choice of an optimal estimator for Θ. Let $F^*(\theta;X)$ denote the mean Fisher information about Θ expected from one observation on X, corresponding to an optimal probability density $f^*(x;\Theta)$. Then choosing a different density function, say $f_1(x;\Theta)$ such that $F_1(\theta;X) > F^*(\theta;X)$, we obtain, by the Cramér-Rao inequality, a sharper lower bound on the variance of an estimator $\tilde{\Theta}$ for Θ, than can be justified with the given prior knowledge. Thus, the accuracy of an optimal estimate $\tilde{\Theta}_1$ based on $F_1(\theta;X)$ will be, as a matter of fact, lower than we might expect. The reason is that $f_1(x;\Theta)$ assumes some non-existing prior knowledge. Suppose $f_2(x;\Theta)$ is another non-optimal density function, such that $F_2(\theta;X) < F^*(\theta;X)$, which results in a larger lower bound on the variance of Θ. The accuracy of an estimate $\tilde{\Theta}_2$ based on $F_2(\theta;X)$ would be lower than could be expected.
This is because in choosing $f_2(x;\Theta)$, we neglect some prior knowledge, which should be used to improve the estimate of Θ.

Let us consider some examples. Suppose Θ is known to be a location parameter for an observed random variable X, satisfying (5.2.29). Suppose we also know the variance of X, given by

$$\int (x - m)^2 \; f(x;\theta)\, dx = \sigma^2, \qquad (6.3.7)$$

where m is the mathematical expectation of X and $f(x;\theta)$ is an unknown probability density of X.

Minimizing the Fisher information given in (5.2.4) with respect to $f(x;\theta)$ under constraints given in (5.2.29) and 6.3.7), we obtain an

optimal probability density (Stam, 1959; Kagan, 1973)

$$f^*(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right], \qquad (6.3.8)$$

which is a Gaussian function with θ as a mathematical expectation and a variance σ^2. Note that the minimal Fisher information is equal to $\frac{1}{\sigma^2}$. Consider a more general case, when θ is also known to be a location parameter for X, but in place of (6.3.7) an s-th absolute central moment of X,

$$\int |x-m|^s \, f(x;\theta)dx = C_s, \; s > 1, \qquad (6.3.9)$$

is given. Note that for s = 2, (6.3.9) reduces to (6.3.7). Minimizing the generalized Fisher information given in (5.2.27) with respect to $f(x;\theta)$ under contraints given in (5.2.29) and (6.3.9) leads to an optimal probability density (Boekee, 1977)

$$f^*(x;\theta) = \frac{s^{\frac{s-1}{s}}}{2\Gamma(\frac{1}{s})C_s^{\frac{1}{s}}} \exp\left[-\frac{|x-\theta|^s}{s\, C_s}\right], \; s > 1, \qquad (6.3.10)$$

called an exponential power distribution. For s = 2, (6.3.10) reduces to the normal distribution given in (6.3.8). The minimal value of a generalized fisher information defined in (5.2.27), corresponding to the optimal function $f^*(x;\theta)$ given in (6.3.10), is equal to

$$F_s^*(\theta;x) = \frac{1}{C_s}.$$

for s = 2, we have $F^*(\theta;x) = \frac{1}{\sigma^2}$, which coincides with the mininal Fisher information in the example above. It should be mentioned that both (6.3.8) and (6.3.10) appear to be the optimal solutions for an unknown (non-parametric) probability density f(x), which can be obtained by the maximum entropy principle under a single constraint given in (6.3.7) or (6.3.9), respectively. (cf. (Boekee, 1977a)).

Suppose now that Θ is known to be a scale parameter for an observed random variable X, satisfying (5.2.30). Suppose further that the prior knowledge on X is given by

$$X > 0, \tag{6.3.11}$$

implying that X is a positive valued variable,

$$\int x\, f(x;\theta)\, dx = m \tag{6.3.12}$$

and

$$\int (x-m)^2\, f(x;\theta)\, dx = \sigma^2, \tag{6.3.13}$$

where m and σ^2 are the known mean and variance of X, respectively.

Minimizing the Fisher information (5.2.4) under constraints given in (5.2.30), (6.3.11) to (6.3.13) results in an optimal probability density (Kagan, 1973)

$$f^*(x;\theta) = \frac{1}{\theta\Gamma(\frac{m}{\theta})} \left[\frac{x}{\theta}\right]^{\frac{m}{\theta}-1} \exp(-\frac{x}{\theta}), \quad x > 0, \tag{6.3.14}$$

which is a gamma distribution with $\theta = \frac{\sigma^2}{m}$.

The minimum information principle can be extended to a random parameter Θ as follows. In this case we have to choose both the probability density $f(x;\Theta)$ for an observed random variable X and the probability density $f(\theta)$ for an unknown parameter Θ. The Fisher information about a random parameter Θ expected from one observation on X is given by (5.2.5), which suggests the following optimization procedure. Making use of the minimum information principle discussed above, we can find an optimal function $f^*(x;\theta)$ assuming Θ to be non-random. Next we determine the minimum Fisher information $F^*(\theta;X)$ given in (5.2.4), corresponding to $f^*(x;\theta)$. Setting $F^*(\theta;X)$ for $F(\theta;X)$ in the right hand side of (5.2.5) and subsequent minimization of the Fisher information $F(\Theta;X)$ given in (5.2.5) with respect to $f(\theta)$ under constraints imposed by a prior knowledge on Θ, will yield an optimal probability density $f^*(\theta)$ sought.

If the function $F^*(\theta;X)$, obtained after the first minimization procedure, appears to be independent of Θ, any probability density $f(\theta)$ will satisfy the optimality condition in the second minimization. This implies that $f(\theta)$ is irrelevant from the point of view of the Fisher information theory. The right hand side in the Cramér-Rao inequality (5.2.28) also appears to be independent of $f(\theta)$ and thus the lower bounds on the variance of the best estimator $\tilde{\theta}$ will be the same for all possible distributions of Θ.

We might, however, need the probability density $f(\theta)$ for other purposes. In that case an optimal $f^*(\theta)$ can be obtained by the maximum entropy principle, as such a density function $f(\theta)$, which yields (under certain constraints imposed by a prior knowledge on Θ) the maximal entropy of Θ given in

$$H(\Theta) = -\int f(\theta) \log f(\theta)\, d\theta. \tag{6.3.15}$$

Note that $F(\theta;X)$ is independent of θ, e.g., if Θ is a location parameter. For a scale parameter, the Fisher information $F(\theta;X)$ can be represented by

$$F(\theta;X) = \frac{1}{\theta^2} F(1;X) \tag{6.3.16}$$

(see section 5.2). The minimal Fisher information corresponding to an optimal probability density $f^*(x;\theta)$ can be expressed by a similar relation

$$F^*(\theta;X) = \frac{1}{\theta^2} F^*(1;X). \tag{6.3.17}$$

where $F^*(1;X)$ denotes the minimal Fisher information corresponding to an optimal $f^*(x;\theta)$, with $\theta = 1$. Setting (6.3.17) in the right hand side of (5.2.5) leads to

$$F^*(\Theta;X) = F^*(1;X) \int \frac{1}{\theta^2} f(\theta)\, d\theta. \tag{6.3.18}$$

Thus, minimizing $F^*(\Theta;X)$ with respect to $f(\theta)$ appears to be equivalent to minimizing

$$E_\Theta \left[\frac{1}{\theta^2}\right] = \int \frac{1}{\theta^2} f(\theta)\, d\theta. \tag{6.3.19}$$

It follows that an optimal density function $f^*(\theta)$ of a scale parameter

Θ can be found by minimizing (6.3.19) with respect to $f(\theta)$ under constraints given by a prior knowledge on Θ.

6.4 Information approach to identification of structure of the model

The information-theoretic approach appears to be useful for determination of the structure (dimension) of a given statistical model. Suppose the model is represented by a probability density $f(x;\theta^k)$, where θ^k is an unknown vector valued parameter $(\theta_1,\ldots,\theta_k)$ with an unknown dimension k. The problem exists in the choice of such a density function $f(x;\tilde{\theta}^{\tilde{k}})$, which would be the best approximation of the true distribution $f(x;\theta^k)$ corresponding to a given observation sample $x = (x_1,\ldots,x_N)$. A solution for an optimal $\tilde{k}$ based on an information-theoretic consideration has been suggested by Akaike (1973; 1981). We shall review this problem by making use of the measure of inaccuracy given in (6.1.14), which seems to be illuminating in this case.

Inaccuracy of the estimated probability density $\tilde{f} = f(x;\tilde{\theta}^{\tilde{k}})$ with respect to the true probability distribution $f = f(x;\theta^k)$ can be expressed, according to (6.1.14), by

$$\bar{A}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = -\int f(x;\theta^k)\ \log\ f(x;\tilde{\theta}^{\tilde{k}})\ dx = E_f[-\log\ f(x;\tilde{\theta}^{\tilde{k}})], \tag{6.4.1}$$

where E_f denotes the mathematical expectation with respect to the true density function $f(x;\theta^k)$. Given a sample $x = (x_1,\ldots,x_N)$, we can estimate this expectation by the sample mean log-likelihood function

$$\tilde{\bar{A}}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = -\frac{1}{N}\sum_{i=1}^{N} \log\ f(x_i;\ \tilde{\theta}^{\tilde{k}}). \tag{6.4.2}$$

Minimization (6.4.2) with respect to $\tilde{\theta}^{\tilde{k}}$ (note that both $\tilde{\theta}_i$, $i=1,\ldots,k$ and $\tilde{k}$ are unknown) will yield the solution sought.

The estimate (6.4.2) is known, however, to be biased. In order to improve $\tilde{\theta}^{\tilde{k}}$, let us return to the relation (6.1.15), which now attains the form

$$\bar{A}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = H(f) + J\begin{bmatrix} f \\ \tilde{f} \end{bmatrix}, \tag{6.4.3}$$

where

$$H(f) = -\int f(x;\theta^k) \log f(x;\theta^k)\, dx \tag{6.4.4}$$

is the Shannon entropy of X corresponding to its true probability distribution and

$$J\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = \int f(x;\theta^k) \log \frac{f(x;\theta^k)}{f(x;\tilde{\theta}^{\tilde{k}})}\, dx \tag{6.4.5}$$

is the Kullback-Leibler divergence of $f(x;\tilde{\theta}^{\tilde{k}})$ from the true density function $f(x;\theta^k)$.

The first term on the right hand side of (6.4.3) is thus determined by the true density function f and can be regarded as a constant. Then minimizing (6.4.2) with respect to $\tilde{\theta}^{\tilde{k}}$ is by (6.4.3) and (6.4.4) equivalent to minimizing the estimate of divergence given in (6.4.5), which is actually an estimate of the log-likelihood ratio,

$$\tilde{J}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{f(x_i;\theta^k)}{f(x_i;\tilde{\theta}^{\tilde{k}})} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{f(x_i;\tilde{\theta}^{\tilde{k}})}{f(x_i;\theta^k)}. \tag{6.4.6}$$

Suppose the true distribution $f(x;\theta^k)$ to be given. In that case the statistic

$$2N\tilde{J}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = -2\sum_{i=1}^{N} \log \frac{f(x_i;\tilde{\theta}^{\tilde{k}})}{f(x_i;\theta^k)} \tag{6.4.7}$$

is known to be asymptotically distributed as χ^2 with $k-\tilde{k}$ degrees of freedom. Thus it follows that the mathematical expectation of the etimate (6.4.6) is equal to

$$E\left[\tilde{J}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix}\right] = \frac{k - \tilde{k}}{2N}. \qquad (6.4.8)$$

Taking into consideration (6.4.3) and (6.4.8), the mathematical expectation of the estimate (6.4.2) may be written as

$$E\left[\tilde{\bar{A}}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix}\right] = H(f) + E\left[\tilde{J}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix}\right] = H(f) + \frac{k - \tilde{k}}{2N}. \qquad (6.4.9)$$

The last term on the right hand side of (6.4.9) is the bias of the estimate of inaccuracy sought. The improved (unbiased) estimate is given by

$$\tilde{\bar{A}}\begin{bmatrix} f \\ \tilde{f} \end{bmatrix} = -\frac{1}{N}\sum_{i=1}^{N} \log f(x_i;\tilde{\theta}^{\tilde{k}}) - \frac{k}{2N} + b(\tilde{k},N), \qquad (6.4.10)$$

where $b(\tilde{k},N) = \frac{\tilde{k}}{2N}$.

Taking into account that the number k, which occurs in (6.4.10), denotes the true dimension given in $f(x;\theta^k)$ and thus may be treated as a constant, we can see that minimizing (6.4.10) with respect to $\tilde{\theta}^{\tilde{k}}$ is equivalent to minimizing

$$W(x;\tilde{\theta}^{\tilde{k}}) = -\sum_{i=1}^{N} \log f(x_i;\tilde{\theta}^{\tilde{k}}) + b_1, \qquad (6.4.11)$$

where $b_1 = \frac{\tilde{k}}{2}$.

It should be mentioned that (6.4.8) and hence the final criterion (6.4.11) have been obtained by assuming the true distribution $f(x;\theta^k)$ to be known. As a matter of fact, however, θ^k is unknown. On account of this, Akaike arrives at a different last term on the right hand side of (6.4.11) given by

$$b_1' = b_1 + \frac{\tilde{k}}{2} = \tilde{k}. \qquad (6.4.12)$$

It appears, however, that both b_1 and b_1' yield inconsistent estimates

$\tilde{k}$: the probability of getting a correct estimate of k differs from unity even if N tends to infinity. Kashyap (1980) has shown that minimizing (6.4.11) with respect to $\tilde{\Theta}^{\tilde{k}}$ can give a consistent estimate for k only if the last term on the right hand side attains the form

$$b_2 = \tilde{k}\ \phi(N), \tag{6.4.13}$$

where ϕ is any function of N such that $\phi(N) > 0$ and if $N \to \infty$, then $\phi(N) \to \infty$ and $\frac{\phi(N)}{N} \to 0$. This seems to be consistent with the result reported by Schwarz (1978), who arrived at

$$b_2' = \frac{\tilde{k}}{2} \log N = \tilde{k} \log \sqrt{N} \tag{6.4.14}$$

(cf. (Kashyap, 1977). The use of $\phi(N) = \ln N$ has also been discussed by Akaike (1977). Asymptotic properties of Akaike's criterion have been studied by Shibata (1976).

A different approach to identification of structure of the model, based on the minimax entropy principle (see section 6.2), has been developed by Rissanen (1976; 1976a). Suppose $X = X(\Theta^n)$ to be an observed variable related to an estimated vector valued parameter $\Theta^n = (\Theta_1,\ldots,\Theta_n)$ with unknown n. The probability densities of both X and Θ^n are assumed to be known. If this is not the case, we can choose the best suitable distributions, corresponding to the prior knowledge available, by the maximum entropy principle discussed in the preceding section. We can take, e.g., normal distributions, provided the variances of X and Θ^n are known.
Let $\Theta^{ik} = (\Theta_{i_1},\ldots,\Theta_{i_k})$, $k \leq n$ denote the true parameter, where $i_1,\ldots i_k$ are different from each other, the numbers taking on values $1,\ldots,n$. Given a sample of observations $x^N = (x_1,\ldots,x_N)$, the best estimate $\tilde{\Theta}^{\tilde{i}\tilde{k}} = (\tilde{\Theta}_{\tilde{i}_1},\ldots,\tilde{\Theta}_{\tilde{i}_{\tilde{k}}})$ of Θ^{ik} is supposed to yield the

minimal joint entropy

$$H(\tilde{\theta}^{\tilde{i}\tilde{k}},\hat{x}^N) = -\int f(\tilde{\theta}^{\tilde{i}\tilde{k}}, \hat{x}^N) \log f(\tilde{\theta}^{\tilde{i}\tilde{k}},\hat{x}^N). \qquad (6.4.15)$$

where $\hat{x}^N = \tilde{x}^N - x^N = (\tilde{x}_1-x_1,\ldots,\tilde{x}_N-x_N)$ is the prediction error, $\tilde{x}^N = (\tilde{x}_1,\ldots,\tilde{x}_N)$ being the estimate of x^N evaluated on the basis of the estimate $\tilde{\theta}^{\tilde{i}\tilde{k}}$, and $f(\tilde{\theta}^{\tilde{i}\tilde{k}},\hat{x}^N)$ denotes the joint probability density of the estimate $\tilde{\theta}^{\tilde{i}\tilde{k}}$ and the prediction error $\hat{x}^N$.

Another requirement imposed upon an optimal $\tilde{\theta}^{\tilde{i}\tilde{k}}$ exists in that its dimension $\tilde{k}$ should be minimal. Assuming, by the maximum entropy principle, the distribution $f(\tilde{\theta}^{\tilde{i}\tilde{k}},\hat{x}^N)$ to be normal, Rissanen has shown that minimizing (6.4.15) is equivalent to minimizing (with respect to i and k)

$$W(\hat{x}^N,\tilde{\theta}^{\tilde{i}\tilde{k}}) = \log\det \tilde{P}^{\tilde{i}\tilde{k}}(\hat{x}^N) + \frac{1}{N}\sum_{j=1}^{\tilde{k}} \log \tilde{\sigma}^2(\tilde{\theta}_{\tilde{i}\tilde{j}}) + \frac{\tilde{k}}{N}(1+\log(2\pi)), \qquad (6.4.16)$$

where $\tilde{P}^{\tilde{i}\tilde{k}}(\hat{x}^N)$ is the prediction error covariance matrix, evaluated with the given sample x^N and the estimate $\tilde{\theta}^{\tilde{i}\tilde{k}}$, and $\tilde{\sigma}^2(\tilde{\theta}_{\tilde{i}\tilde{i}})$ is the sample estimate of the variance of $\tilde{\theta}_{\tilde{i}\tilde{j}}$ (Rissanen, 1976). As compared with Akaike's information criterion given in (6.4.11) together with (6.4.12) Rissanen's criterion (6.4.16) tends to yield a lower order $\tilde{k}$ of the estimated model.

Ishii and Suzumura (1977) have used an information-theoretic approach for estimating the order of an autoregressive process. Suppose the process to be given by a sequence $X_1,\ldots,X_m$. The conditional entropy of X_k given $X_{k-1},\ldots,X_1$ can be expressed as

$$H_k = H(X_k/X_{k-1},\ldots,X_1) = -\int f(x_1,\ldots,x_k)\log f(x_k/x_{k-1},\ldots,x_1)dx_1\ldots dx_k, \quad k = 1,\ldots,m, \qquad (6.4.17)$$

where $f(x_1,\ldots,x_k)$ is the joint probability density of $X_1,\ldots,X_k$ and $f(x_k/x_{k-1},\ldots,x_1)$ is the conditional probability density of X_k given $X_{k-1},\ldots,X_1$. The optimal estimate of the order of the

given process can be found as such a number $\tilde{k}$, which satisfies the following relation

$$H_1 > H_2 > \ldots > H_{\tilde{k}} = H_{\tilde{k}+1} = \ldots \quad . \tag{6.4.18}$$

An information measure of the degree of identity for statistical models has been introduced by Durgaryan and Rajbman (1968; see also (Rajbman, 1980)). This measure is given by

$$q(X,Y) = \frac{I(Y;X)}{H(Y)}, \tag{6.4.19}$$

where H(Y) is the Shannon entropy of the model output variable Y and I(Y;X) is the Shannon information about the input variable X, provided by Y. The quantity (6.4.19) seems to give a suitable criterion for discrimination between the competing models, including multi-input multi-output ones.

An extended Kullback-Leibler divergence, satisfying the triangle inequality and thus being a metric measure, due to Baram and Sandell (1977) appears to be an effective criterion for estimating the order of dynamic processes.

For a list of publications on information theory and identification, see (Ponomarenko, 1981).

7. CONCLUSIONS

Identification is a cognitive process whose principle task is the acquisition of knowledge about an object. The information approach thus seems to be a good basis for discussing methods/techniques, for deriving objective criteria and for presenting a unified picture of identification in general.

The most natural information model of identification is suggested by the well-established Shannon information theory, which considers

information as a reduction in uncertainty about the estimated phenomena due to observations. The Shannon information measure is based on the prior knowledge of the joint probability distribution for both observed and estimated variables. Different forms of the prior knowledge suggest alternative information models. Certain generalizations of the Shannon information measure such as information of order α or information of type β, as well as other information models dicussed in the present report may appear more appropriate in specific identification problems, dependent on the form of the prior knowledge and on its kind of statistical representation.

An intuitively appealing information model results, for example, by making use of the concept of inaccuracy introduced by Kerridge. The concept of certainty due to Van der Lubbe, which can be regarded as an analogue of the concept of information, leads to another information model of identification, since the latter is aimed at raising the degree of certainty concerning the estimated object.

From the point of view of identificiation, the main distinction between different information models lies in the form of the probability distributions involved. Parametric and non-parametric representations result in different information models. The relations between several information measures discussed in chapter 6 seem to be useful for a better understanding of the respective information concepts as well as for the choice of a suitable information model in a given identification problem.

Information theory provides the most general approach to identification, suggesting optimal solutions for many problems. The maximum entropy principle, together with its extension termed minimum information principle, appear to be the only tools available so far, for the choice of the prior probability distribution related to an estimated process. Optimal information-theoretic solutions can also be derived in such problems as the choice of the best input test signals or determination of structure (order) of the model. The basic information-theoretic estimation principle suggests an appropriate approach to

traditional problems in estimation (prediction, filtering, smoothing) and parameter identification.

Considerations presented in this report indicate that information theory is a good basis for generalization, unification and further development in the field of estimation and identification.

REFERENCES

Aczél, J. and Z. Daróczy (1963)
Charakterisierung der Entropien positiver Ordnung und der Shannonschen Entropie.
Acta Math. Acad. Sci. Hungar., Vol. 14, p. 95-121.

Aczél, J. and Z. Daróczy (1975)
On measures of information and their characterizations.
New York: Academic Press.
Mathematics in Science and Engineering, Vol. 115.

Aggarawal, N.L. (1974)
Sur l'information de Fisher.
In: Théories de l'information. Actes des Rencontres de Marseille-Luminy, 5 au 7 juin 1973. Ed. by J. Kampé de Fériet and C.F. Picard.
Berlin: Springer 1974.
Lecture Notes in Mathematics, Vol. 398. P. 111-117.

Akaike, H. (1973)
Information theory and an extension of the maximum likelihood principle.
In: Proc. 2nd Int. Symp. on Information Theory; Tsahkadsor, Armenia, 2-8 Sept. 1971. Ed. by B.N. Petrov and F. Csáki.
Budapest: Akadémiai Kiadó. P. 267-281.

Akaike, H. (1977)
On entropy maximization principle.
In: Applications of Statistics. Proc. Symp.; Dayton, 14-18 June 1976. Ed. by P.R. Krishnaiah.
Amsterdam: North Holland. P. 27-41.

Akaike, H. (1981)
Modern development of statistical methods.
In: Trends and Progress in System Identification. Ed. by P. Eykhoff.
Oxford: Pergamon.
IFAC Series for Graduates, Research Workers & Practising Engineers,
Vol. 1. P. 169-184.

Arimoto, S. (1971)
Information-theoretical considerations on estimation problems.
Inf. & Control, Vol. 19, p. 181-194.

Arimoto, S. (1975)
Information measures and capacity of order α for discrete memoryless channels.
In: Topics in information theory. Proc. 2nd Colloquium on Information Theory; Keszthely, Hungary, 25-29 Aug. 1975.
Ed. by I. Csiszár and P. Elias.
Amsterdam: North-Holland, 1977.
Colloquia Mathematica Societatis János Bolyai, Vol. 16. r. 41-52.

Arimoto, S. and H. Kimura (1971)
Optimum input test signals for system identification - an information-theoretical approach.
Int. J. Syst. Sci., Vol. 1, p. 279-290.

Baram, Y. and N.R. Sandell Jr. (1977)
An information theoretic approach to dynamical systems modeling and identification.
In: Proc. 1977 IEEE Conf. on Decision & Control. Including the 16th Symp. on Adaptive Processes and a Special Symp. on Fuzzy Set Theory and Applications; New Orleans, 7-9 Dec. 1977.
New York: IEEE. P. 1113-1118.

Barankin, E.W. (1949)
Locally best unbiased estimates.
Ann. Math. Stat., Vol. 20, p. 477-501.

Ben-Bassat, M. and J. Raviv (1978)
Rényi's entropy and the probability of error.
IEEE Trans. Inf. Theory, Vol. IT-24, p. 324-331.

Bhattacharyya, A. (1943)
On a measure of divergence between two statistical populations defined by their probability distributions.
Calcutta Mathematical Society Bulletin, Vol. 35, p. 99-109.

Bhattacharyya, A. (1946)
On some analogues of the amount of information and their use in statistical estimation.
Sankhya, Vol. 8, p. 1-14, 201-218.

Boekee, D.E. (1975)
An extension of the Fisher information measure.
In: Topics in information theory. Proc. 2nd Colloquium on Information Theory; Keszthely, Hungary, 25-29 Aug. 1975.
Ed. by I. Csiszár and P. Elias.
Amsterdam: North-Holland, 1977.
Colloquia Mathematica Societatis János Bolyai, Vol. 16. P. 113-123.

Boekee, D.E. (1976)
Maximum information in continuous systems with contstraints.
In: Proc. 8th Int. Congress on Cybernetics; Namur, 6-11 Sept. 1976.
Namur: Association Internationale de Cybernétique, 1977. P. 243-259.

Boekee, D.E. (1977)
A generalization of the Fisher information measure.
Dr. Thesis. Delft University of Technology, Netherlands.
Delft University Press.

Boekee, D.E. (1978)
Generalized Fisher information with application to estimation problems.
In: Information and Systems. Proc. IFAC Workshop, Compiegne, France, 25-27 Oct. 1977. Ed. by B. Dubuisson.
Oxford: Pergamon, 1978. P. 75-82.

Boekee, D.E. and J.C.A. van der Lubbe (1980)
The R-norm information measure.
Inf. & Control, Vol. 45, p. 136-155.

Chapman, D.G. and H. Robbins (1951)
Minimum variance estimation without regularity assumptions.
Annals of Mathematical Statistics, Vol. 22, p. 581-586.

Chaundy, T.W. and J.B. McLeod (1960)
On functional equation.
Edinburgh Math. Notes, Vol. 43, p. 8.

Csiszár, I. (1967)
Information-type measures of difference of probability distributions and indirect observations.
Studia Sci. Math. Hungar., Vol 2, p. 299-318.

Csiszár, I. (1978)
Information measures: a critical survey.
In: Trans. 7th Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes and the European Meeting of Statisticians; Prague, 18-23 Aug. 1974. Vol. B.
Dordrecht: Reidel, 1978. P. 73-86.

Daróczy, Z. (1964)
Über Mittelwerte und Entropien vollständiger Wahrscheinlichkeitsverteilungen.
Acta Math. Acad. Sci. Hungar., Vol. 15, p. 203-210.

Daróczy, Z. (1969)
On the Shannon measure of information (Hungarian).
Magyar Tod. Akad. Mat. Fiz. Oszt. Közl., Vol. 19, p. 9-24.
English Transl.: Selected Translations in Mathematical Statistics and Probability, 1972, Vol. 10, p. 193-210.

Daróczy, Z. (1970)
Generalized information functions.
Inf. & Control, Vol. 16, p. 36-51.

Durgaryan, I.S. and N.S. Rajbman (1968)
An information measure of isomorphism of an object and its model as well as its application to automatic control. (Russian).
Measuring and Monitoring Systems. All-Union Seminar on Information Methods in Control, Vladivostok.

Eykhoff, P. (1980)
System identification: approach to a coherent picture through template functions.
Electron. Lett., Vol. 16, p. 502-504.

Eykhoff, P., A.J.W. van den Boom and A.A. van Rede (1981)
System identification methods: - unification and information development using template functions.
In: Control Science and Technology for the Progress of Society. Preprints 8th IFAC World Congress; Kyoto, 24-28 Aug. 1981. Vol. 6. P. 83-88.

Fadeev, D.K. (1956)
On the concept of entropy of a finite probabilistic scheme (Russian).
Uspekhi Mat. Nauk, Vol. 11, no. 1 (67), p. 227-231.

Fisher, R.A. (1925)
Theory of statistical estimation.
Proceedings of the Cambridge Philosophical Society, Vol. 22, p. 700-725.

Forte, B. and C.T. Ng (1973)
On a characterization of the entropies of degree β.
Utilitas Mathematica, Vol. 4, p. 193-205.

Gart, J.J. (1959)
An extension of the Cramér-Rao inequality.
Ann. Math. Stat., Vol. 30, p. 367-380.

Havrda, J. and F. Charvát (1967)
Quantification method of classification process. Concept of structural α-entropy.
Kybernetika, Vol. 3, p. 30-35.

Ishii, N. and N. Suzumura (1977)
Estimation of the order of autoregressive process.
Int. J. Syst. Sci., Vol. 8, p. 905-913.

Jaynes, E.T. (1968)
Prior probabilities.
IEEE Trans. Syst. Sci. & Cybern., Vol. SSC-4, p. 227-241.

Kagan, A.M. (1963)
On the theory of Fisher's amount of information.
Sov. Math.-Doklady, Vol. 4, p. 991-993.
(Transl. of "Doklady Akademii Nauk SSSR", Vol. 151, &963, No. 1-6).

Kagan, A.M., Y.V. Linnik and C.R. Rao (1973)
Characterization problems in mathematical statistics.
New York: Wiley.
Wiley Series in Probability and Mathematical Statistics.

Kalata, P. and R. Priemer (1974)
On minimal error entropy stochastic approximation.
Int. J. Syst. Sci., Vol. 5, p. 895-906.

Kalata, P. and R. Priemer (1978)
On system identification with and without certainty.
J. Cybern., Vol. 8, No. 1, p. 31-50.

Kalata, P. and R. Priemer (1979)
Linear prediction, filtering, and smoothing: an information-theoretic approach.
Inf. Sci., Vol. 17, p. 1-14.

Kannappan, Pl. (1972)
On directed divergence and inaccuracy.
Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, Vol. 22, p. 49-55.

Kannappan, Pl. (1972a)
On Shannon's entropy, directed divergence and inaccuracy.
Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, Vol. 22, p. 95-100.

Kannappan, Pl. and P.N. Rathie (1973)
On a characterization of directed divergence.
Inf. & Control, Vol. 22, p. 163-171.

Kapur, J.N. (1968)
Information of order α and type β.
Proc. Indian Acad. Sci., Vol. 68, p. 65-75.

Kashyap, R.L. (1977)
A Bayesian comparison of different classes of dynamic models using empirical data.
IEEE Trans. on Automatic Control, Vol. AC-22, p. 715-727.

Kashyap, R.L. (1980)
Inconsistency of the AIC rule for estimating the order of auto-regressive models.
IEEE Trans. on Automatic Control, Vol. AC-25, p. 996-998.

Kerridge, D.F. (1961)
Inaccuracy and inference.
J. Royal Stat. Soc. Ser. B, Vol. 23, p. 184-194.

Khinchin, A.I. (1953)
The entropy concept in probability theory (Russian).
Uspekhi Mat. Nauk, Vol. 8, no. 3(55), p. 3-20.
English Transl.: Mathematical Foundations of Information Theory, 1957, Dover, New York. P. 1-28.

Kopilovich, L. Ye. (1964)
Distribution laws of an envelope of random processes with maximum entropy.
Radio Eng. & electron. Phys., Vol. 9, p. 268-272.
(Transl. of "Radiotekh. & Elektron.").

Kullback, S. (1959)
Information theory and statistics
New York: Wiley.
Wiley Publications in Statistics.

Kullback, S. and R.A. Leibler (1951)
On information and sufficiency.
Ann. Math. Stat., vol. 22, p. 79-86.

Luce, R.D. (1960)
The theory of selective information and some of its behaviorial applications.
In: Developments in Mathematical Psychology, Information, Learning and Tracking, Free Press of Glencoe, Glencoe, Illinois. P. 1-119.

Mathai, A.M. and P.N. Rathie (1975)
Basic concepts in information theory and statistics: Axiomatic foundations and applications.
New Delhi: Wiley Eastern.

Noonan, J.P., N.S. Tzannes and T. Costello (1976)
On the inverse problem of entropy and maximizations.
IEEE Trans. Inf. Theory, Vol. IT-22, p. 120-123.

Papaioannou, P.C. (1970)
On statistical information theory and related measures of information.
Ph.D. Thesis. Iowa State University of Science and Technology, Ames.
Available from: University Microfilms, Ann Arbor, Mich., USA.
Order No. 70-25815.

Patni, G.C. and K.C. Jain (1976)
On some information measures.
Inf. & Control, Vol. 31, p. 185-192.

Ponomarenko, M.F. (1981)
Information measures and their applications to identification (a bibliography).
Eindhoven University of Technology, Department of Electrical Engineering.
Eindhoven University of Technology Research Reports, EUT Report 81-E-123.

Rajbman, N.S. and V.M. Chadeev (1980)
Identification of industrial processes: The application of computers in research and production control.
Amsterdam: North Holland, 1980.
(Russian ed.: Energiya, Moscow, 1975).

Rao, C.R. (1945)
Information and the accuracy attainable in the estimation of statistical parameters.
Bull. Calcutta Math. Soc., Vol. 37, p. 81-91.

Rathie, P.N. (1970)
On generalized measures of inaccuracies, information and errors in information.
Statistica, Vol. 30, p. 340-349.

Rathie, P.N. (1971)
On some new measures of uncertainty, inaccuracy and information and their characterizations.
Kybernetika, Vol. 7, p. 394-403.

Rathie, P.N. and Pl. Kannappan (1972)
A directed-divergence function of type β.
Inf. & Control, Vol. 20, p. 38-45.

Rathie, P.N. and Pl. Kannappan (1973)
An inaccuracy function of type β.
Ann. Inst. Statist. Math., Vol. 24, p. 205-214.

Rathie, P.N. and P. Nath (1972)
On inaccuracies, β-inaccuracies and errors in information.
Univ. Nac. Tucuman Rev., Ser.A, Vol. 22, p. 7-16.
(Universidad nacional de Tucumán, San Miguel de Tucumán, Argentina).

Rényi, A. (1961)
On measures of entropy and information.
In: Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability. Vol. 1: Contributions to the Theory of Statistics.
Berkeley, 20 June - 30 July 1960. Ed. by J. Neyman.
Berkeley - Los Angeles: University of California Press, 1961.
P. 547-561.

Rissanen, J. (1976)
Minimax entropy estimation of models for vector processes.
In: System Identification: Advances and case studies.
Ed. by R.K. Mehra and D.G. Lainiotis.
New York: Academic Press.
Mathematics in Science and Engineering, Vol. 126. P. 97-119.

Rissanen, J. (1976a)
Parameter estimation by shortest description of data.
In: Proc. 16th Joint Automatic Control Conf.; West Lafayette, 27-30 July 1976.
New York: American Society of Mechanical Engineers, 1976. P. 593-597.

Rozenberg, V. Ya. and N.A. Rubichev (1966)
An inverse problem in information theory.
Probl. Inf. Transm., Vol. 2, No. 2, p. 63-64.
(Transl. of "Probl. Peredachi Inf.").

Schwarz, G. (1978)
Estimating the dimension of a model.
Ann. Stat., Vol. 6, p. 461-464.

Shannon, C.E. (1948)
A mathematical theory of communication.
Bell Syst. Techn. J., Vol. 27, p. 379-423; 623-656.

Sharma, B.D. and R. Autar (1973)
On characterization of a generalized inaccuracy measure in information theory.
J. Appl. Probab., Vol. 10, p. 464-468.

Sharma, B.D. and R. Autar (1973)
Relative-information functions and their type (α,β) generalizations.
Metrika, Vol. 21, p. 41-50.

Sharma, B.D. and D.P. Mittal (1975)
New non-additive measures of entropy for discrete probability distributions.
J. Math. Sci., Vol. 10, p. 28-40.

Shibata, R. (1976)
Selection of the order of an autoregressive model by Akaike's information criterion.
Biometrika, Vol. 63, p. 117-126.

Stam, A.J. (1959)
Some mathematical properties of quantities of information.
Dr. Thesis, Delft University of Technology, Delft, Netherlands.

Tomita, Y., S. Ohmatu and T. Soeda (1976)
An application of the information theory to filtering problems.
Inf. Sciences, Vol. 11, p. 13-27.

Tomita, Y., S. Ohamatsu and T. Soeda (1976a)
An application of the information theory to estimation problems.
Inf. & Control, Vol. 32, p. 101-111.

Tverberg, H. (1958)
A new derivation of the information function.
Math. Scand., Vol. 6, p. 297-298.

Vajda, I. (1973)
χ^{α}-divergence and generalized Fisher information.
In: Trans. 6th Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes; Prague, 19-25 Sept. 1971.
Prague: Academia, 1973. P. 873-886.

Van der Lubbe, J.C.A. (1981)
A generalized probabilistic theory of the measurement of certainty and information.
Dr. Thesis. Delft University of Technology, Netherlands. Department of Electrical Engineering, Delft University of Technology.
Technical Report IT-81-02.

Van der Lubbe, J.C.A. and D.E. Boekee (1977)
R-norm information.
Proc. 10th European Meeting of Statisticians; Leuven, Belgium, 22-26 Aug. 1977. P. 173.

Weidemann, H.L. (1969)
Entropy analysis of feedback control systems.
In: Advances in Control Systems. Vol. 7. Ed. by C.T. Leondes.
New York: Academic Press. P. 225-255.

Weidemann H.L. and E.B. Stear (1970)
Entropy analysis of estimating systems.
IEEE Trans. Inf. Theory, Vol. IT-16, p. 264-270.

Wiener, N. (1948)
Cybernetics or Control and Communication in the Animal and the Machine.
Cambridge: MIT Press (1948); MIT Press & Wiley, 2nd ed. (1961).

EINDHOVEN UNIVERSITY OF TECHNOLOGY
THE NETHERLANDS
DEPARTMENT OF ELECTRICAL ENGINEERING

Reports:

105) Videc, M.F.
STRALINGSVERSCHIJNSELEN IN PLASMA'S EN BEWEGENDE MEDIA: Een geometrisch-optische en een golfzonebenadering.
TH-Report 80-E-105. 1980. ISBN 90-6144-105-6

106) Hajdasiński, A.K.
LINEAR MULTIVARIABLE SYSTEMS: Preliminary problems in mathematical description, modelling and identification.
TH-Report 80-E-106. 1980. ISBN 90-6144-106-4

107) Heuvel, W.M.C. van den
CURRENT CHOPPING IN SF_6.
TH-Report 80-E-107. 1980. ISBN 90-6144-107-2

108) Etten, W.C. van and T.M. Lammers
TRANSMISSION OF FM-MODULATED AUDIOSIGNALS IN THE 87.5 - 108 MHz BROADCAST BAND OVER A FIBER OPTIC SYSTEM.
TH-Report 80-E-108. 1980. ISBN 90-6144-108-0

109) Krause, J.C.
SHORT-CURRENT LIMITERS: Literature survey 1973-1979.
TH-Report 80-E-109. 1980. ISBN 90-6144-109-9

110) Matacz, J.S.
UNTERSUCHUNGEN AN GYRATORFILTERSCHALTUNGEN.
TH-Report 80-E-110. 1980. ISBN 90-6144-110-2

111) Otten, R.H.J.M.
STRUCTURED LAYOUT DESIGN.
TH-Report 80-E-111. 1981. ISBN 90-6144-111-0

112) Worm, S.C.J.
OPTIMIZATION OF SOME APERTURE ANTENNA PERFORMANCE INDICES WITH AND WITHOUT PATTERN CONSTRAINTS.
TH-Report 80-E-112. 1980. ISBN 90-6144-112-9

113) Theeuwen, J.F.M. en J.A.G. Jess
EEN INTERACTIEF FUNCTIONEEL ONTWERPSYSTEEM VOOR ELEKTRONISCHE SCHAKELINGEN.
TH-Report 80-E-113. 1980. ISBN 90-6144-113-7

114) Lammers, T.M. en J.L. Manders
EEN DIGITAAL AUDIO-DISTRIBUTIESYSTEEM VOOR 31 STEREOKANALEN VIA GLASVEZEL.
TH-Report 80-E-114. 1980. ISBN 90-6144-114-5

115) Vinck, A.J., A.C.M. Oerlemans and T.G.J.A. Martens
TWO APPLICATIONS OF A CLASS OF CONVOLUTIONAL CODES WITH REDUCED DECODER COMPLEXITY.
TH-Report 80-E-115. 1980. ISBN 90-6144-115-3

EINDHOVEN UNIVERSITY OF TECHNOLOGY
THE NETHERLANDS
DEPARTMENT OF ELECTRICAL ENGINEERING

Reports:

EUT Reports are a continuation of TH-Reports.

116) Versnel, W.
THE CIRCULAR HALL PLATE: Approximation of the geometrical correction factor for small contacts.
TH-Report 81-E-116. 1981. ISBN 90-6144-116-1

117) Fabian, K.
DESIGN AND IMPLEMENTATION OF A CENTRAL INSTRUCTION PROCESSOR WITH A MULTIMASTER BUS INTERFACE.
TH-Report 81-E-117. 1981. ISBN 90-6144-117-X

118) Wang Yen Ping
ENCODING MOVING PICTURE BY USING ADAPTIVE STRAIGHT LINE APPROXIMATION.
EUT Report 81-E-118. 1981. ISBN 90-6144-118-8

119) Heijnen, C.J.H., H.A. Jansen, J.F.G.J. Olijslagers and W. Versnel
FABRICATION OF PLANAR SEMICONDUCTOR DIODES, AN EDUCATIONAL LABORATORY EXPERIMENT.
EUT Report 81-E-119. 1981. ISBN 90-6144-119-6.

120) Piecha, J.
DESCRIPTION AND IMPLEMENTATION OF A SINGLE BOARD COMPUTER FOR INDUSTRIAL CONTROL.
EUT Report 81-E-120. 1981. ISBN 90-6144-120-X

121) Plasman, J.L.C. and C.M.M. Timmers
DIRECT MEASUREMENT OF BLOOD PRESSURE BY LIQUID-FILLED CATHETER MANOMETER SYSTEMS.
EUT Report 81-E-121. 1981. ISBN 90-6144-121-8

122) Ponomarenko, M.F.
INFORMATION THEORY AND IDENTIFICATION.
EUT Report 81-E-122. 1981. ISBN 90-6144-122-6

123) Ponomarenko, M.F.
INFORMATION MEASURES AND THEIR APPLICATIONS TO IDENTIFICATION (a bibliography).
EUT Report 81-E-123. 1981. ISBN 90-6144-123-4

124) Borghi, C.A., A. Veefkind and J.M. Wetzer
EFFECT OF RADIATION AND NON-MAXWELLIAN ELECTRON DISTRIBUTION ON RELAXATION PROCESSES IN AN ATHMOSPHERIC CESIUM SEEDED ARGON PLASMA.
EUT Report 82-E-124. 1982. ISBN 90-6144-124-2

125) Saranummi, N.
DETECTION OF TRENDS IN LONG TERM RECORDINGS OF CARDIOVASCULAR SIGNALS.
EUT Report 82-E-125. 1982. ISBN 90-6144-125-0