

A COMPENDIUM TO INFORMATION THEORY  
IN ECONOMICS AND ECONOMETRICS

By Esfandiar Maasoumi<sup>1</sup>

Department of Economics, SMU, Dallas, TX 75275

Key Words: Information theory, entropy, inequality, tests, adaptive estimation, MLE, distance functions, uncertainty, aggregation, nonparametrics.

JEL classification nos. C13, C14, C50.

ABSTRACT

An extensive synthesis is provided of the concepts, measures and techniques of Information Theory (IT). After an axiomatic description of the basic definitions of "information functions", "entropy" or uncertainty, and the maximum entropy principle, the paper demonstrates the power of IT as both an interpretive and technically productive tool. It is argued that this power and universality is primarily due to the common need for (i) measures of distance and discrimination and, (ii) appropriate partitioning-aggregation properties. IT offers a very suggestive unification for a bewildering and arbitrary set of approaches that have evolved in different disciplines.

Applications are discussed or indicated. These applications have relevance to economics, finance, industrial organization, marketing, statistical inference and model selection, political science and communication. A main focus of the discussion is the generative power of IT measures in statistical examinations of unknown distributions and random phenomena. Measures of concentration and inequality, aggregation functions and index numbers, tests of nested and non-nested hypotheses, and measures of volatility, mobility and divergence are presented. Extending the author's previous work, estimation of unknown regression functions, densities and score functions is examined based on the maximum entropy principle. Some empirical examples are cited.

First draft 1/1992  
Last draft 1/93

---

<sup>1</sup>This paper is an extension and update to my previous surveys in this area, Maasoumi (1988b, 1990). For comments and discussions I thank J. Foster, A. Ullah, reviewers and seminar participants at UC Santa Barbara, UC Riverside, UC San Diego, SMU, Guelph, ESEM Brussels (Aug. 92), Houston and Rice.



## Table of Contents

1. Introduction
  2. Information functionals, expected information and entropy (uncertainty), distance and divergence, axiom systems.
  3. Maximum Entropy (ME) principle and distributions, relation to ML, Min  $\chi^2$ , LS principles, and pseudo-true models.
  4. ME regression functions, Adaptive and Robust Estimation, Tests, Volatility measures.
  5. Akaike information criterion, Minimum Description Length (MDL), other fit and test criteria, Bayesian learning. Fisher's information matrix,  $\varphi$ -entropy matrix, Riemannian geometry and metric spaces.
  6. Axiom systems of inequality and concentration.
  7. Aggregation and multidimensional measures of well-being, mobility and inequality.
  8. Some empirical applications.
- References

## 1. Introduction

The interest in the "distance" between distributions, in its many manifestations, is adopted as a running theme in this paper. Discrimination, tests of hypotheses, goodness of fit and model selection, estimation (MLE, Min  $\chi^2$ , LS, ...), aggregation, clustering and indexing, concentration, inequality, spaces and their metrics, divergence, Maximum Entropy distributions and functions, are all examples of analyses that fall within the domain of Information Theory (IT). The notion of "distance" or "divergence" between distributions, in particular, has been central in statistical inference from the earliest stages. This is evident in the work of Mahalanobis, S. Kullback and Leibler, H. Jeffreys, Pitman, J. Aczel, C.R. Rao, E. Parzen, H. Akaike, J. Rissanen, Shannon, Hartley, A. Renyi, P. Jaynes and many other scholars. In econometrics, G. Tintner, H. Theil, T. Sawa, E. Maasoumi, D. Fiebig, P.M. Robinson, J. Sengupta, H. White, R. Klein, A. Zellner are but a few contributors. The earliest textbook in econometrics to introduce information theory was Davis (1941). Theil's (1967) seminal book and his Principles text in econometrics made IT better known to economists and other social scientists. Anyone who has addressed audiences in economics, however, would attest to the need to make this area better known and appreciated. An account that may achieve some balance between the technical, the

conceptual and the practical aspects of IT may serve a much broader audience. This is our aim in this survey which will have at least as many omissions as any single full length book in this area. A recent text by Cover and Thomas (1991), while written for experts outside of social sciences, provides an excellent coverage of the concepts as well as the applications of information theory in such areas as data compression and communication. An account of IT appeared in Maasoumi (1988b). The current paper is intended to be useful to social scientists and includes some recent developments in economics as well as in econometrics and statistics.

Parzen, moved by the elegance and the power of information theory formalism in a variety of situations, calls for "a new culture in statistics" in which this theory is employed to interpret and motivate most of our activities in statistical inference. In addition, the axiom systems in information theory suggest decomposition principles which distinguish different information functions and "entropies", and identify desirable measures, decision criteria and indices. It is seen that in the absence of these plausible decomposition requirements it would be difficult to justify the currently almost exclusive dependence in applications on Shannon's entropy and the Kullback-Leibler criterion. This analysis suggests new criteria and solutions which open up new directions in research.

In the next section we briefly develop the basic definitions and concepts both for self-sufficiency and in order to provide a flavor of the axiom systems that underly the power and richness of IT as a unifying theory. In subsequent sections I give a discussion of further developments of the modest suggestions in Maasoumi (1979, 1985, 1986b, 1988b) for a maximum entropy approach to non-parametric estimation of unknown regression and aggregation functions. Recent work of Zellner and Highfield (1988) and Maasoumi and Zhang (1992b) on numerical calculation of the Maximum Entropy (ME) density opens the way for practical applications to a host of other functionals such as higher moments, derivatives, score functions, and information matrices. This is a powerful alternative to kernel density estimation which is interpretable as a disciplined method of moments approach.

## 2. Information functions and Axiom Systems

**2.1 Basics:** Let  $p_i = \text{prob}(x=x_i)$ ,  $i=1,2,\dots,n$ . Let an experiment produce an outcome (event)  $x_k$ . What amount of information is conveyed (uncertainty reduced) by this experiment? Following statistical mechanics it has become customary to consider information as a function of probability<sup>2</sup>. Let  $g(p)$  represent the information function. The higher the probability  $p_k$ , before the experiment, the less "news" (order) is conveyed by the outcome. For example, if  $p_k = 1$ , for some  $k$ , and  $p_i = 0$ , for all other  $i \neq k$ , the occurrence of  $x_k$  contains no information. It is then reasonable to require that  $g(\cdot)$  be a decreasing function and  $g(1) = 0$  and  $g(0) \rightarrow \infty$ . There are many non-negative functions that satisfy these requirements.

Axiom (property) systems have been developed in this literature in order to characterize and derive "ideal" information functions. For instance, from Hartley (1928), Erdos (1946) or Renyi (1961) a characterization of  $g(p) = \log(1/p)$  is obtained. The functional analytic results that guide such characterizations are typified by the following basic lemma:

**Lemma.** Let  $h(n)$  be an additive function defined for  $n = 1,2,\dots$ . If we require:

1.  $h(nm) = h(n) + h(m)$ ; (additivity)
2.  $\lim_{n \rightarrow \infty} [h(n+1) - h(n)] = 0$ ; (monotone decreasing)

then  $h(n) = c \ln n$ , where  $c$  is a constant obtained according to the chosen base.

Axiom systems with less stringent "additivity" restrictions lead to generalized information functions such as  $\frac{1-p^{-\gamma}}{\gamma}$  which includes  $-\log p$  as a limiting case when  $\gamma =$

0. We explore axiom systems that characterize the related concepts of expected information and entropy.

### 2.2 Entropy and Expected information.

The expected information from an experiment with  $n$  possible outcomes is defined as follows:

<sup>2</sup>Dependence on probability alone is a convention that is not without controversy; see Georgescu-Roegen (1966).

$$\sum_{i=1}^n p_i g(p_i) = H(p) \geq 0, \quad p = (p_1, p_2, \dots, p_n) \quad (1)$$

When  $g(p_i) = -\log p_i$ ,  $H(p)$  is known as the Shannon (1948) entropy introduced in communication theory. Wiener (1948) proposed the same definition in cybernetics. Also see Shannon and Weaver (1949). For continuous variables the definition in (1) and others below would be extended in the obvious manner using integrals and densities.

Entropy is a measure of uncertainty, disorder or **volatility** associated with a distribution/random variable. For example,  $H(p) = 0$ , if  $p = (0, 0, 1, 0, \dots, 0)$ , and  $H(p) = \text{Max } H(p) = \log n$ , when  $p_i = 1/n$ , all  $i$ . This concept of entropy is inherited from statistical mechanics. It has a controversial and somewhat tenuous connection with the physical concept of "entropy" first introduced as the second law of thermodynamics. The latter posits that the "entropy" of the universe always tends to a maximum ("heat death"). If we think of energy as having two components, active (utilized) and potential (latent), then:

$$\text{Entropy} = \text{potential energy} / \text{absolute temperature}$$

This is an evolutionary law of dissipation. In 1865 R. Clausius coined the word entropy from a greek word for evolution. The statistical concept of entropy arose out of a desire to (i) describe the phenomena (shuffling of particles) that produce heat, and (ii) use probability to represent the inherent disorder in such phenomena so as to derive a definition of entropy (disorder) equivalent to the one above. The principles that allow for this operation to produce, for example, the Boltzmann entropy ( $c \log n$ ), have remained as controversial as the arguments surrounding the definition of probability, subjective or otherwise. It seems best to adopt the definition of statistical entropy as a useful measure of disorder, and to avoid endowment of this definition with a general physical meaning. See Georgescu-Roegen (1966) for an inspiring account.

But statistical entropy is not uniquely defined. There are axiom systems that may justify Shannon's or other entropies. Consider the following axioms:

A1.  $H(p)$  is symmetric (anonymous!)

- A2.  $H(p)$  is continuous in  $p$  for  $0 \leq p_i \leq 1$ .  
 A3.  $H(1/n, 1/n, \dots, 1/n) = 1$ , (or some other normalization).  
 A4.  $H(p * q) = H(p) + H(q)$

where  $q = (q_1, q_2, \dots, q_n)$  and  $p * q$  is the direct product of  $p$  and  $q$  ( $p_i q_i$ ). A4 says that entropies of independent experiments are additive. Under A1 – A4 we can find many entropy functionals; for instance, all linear transformations of the following (Renyi):

$$H_\gamma(p) = \frac{1}{1-\gamma} \log \left( \sum_i p_i^\gamma \right), \quad \gamma \neq 0, 1 \quad (2)$$

and  $\lim_{\gamma \rightarrow 1} H_\gamma(p) = H_1(p) =$  Shannon's entropy defined above.

Now consider the following branching/aggregation axiom in place of A4:

$$\begin{aligned} \text{A5. } & H(p_1, p_2, \dots, p_{k-1}, t p_k, (1-t)p_k, p_{k+1}, \dots, p_n) \\ &= H(p_1, p_2, \dots, p_n) + p_k H(t, 1-t) \text{ for } 0 \leq t \leq 1. \end{aligned}$$

A5 implies A4, but not vice versa. Then:

**Theorem 1.** (Faddeev (1957)):

A1–A3 and A5 identify Shannon's entropy (uniquely).

Compare the following:

**Theorem 2.** (Renyi (1961))

Let  $p \cup q = (p_1, p_2, \dots, p_n, q_1, q_2, \dots, q_m)$  and  $w(p) = \sum_{i=1}^n p_i$ ,  $w(q) = \sum_{i=1}^m q_i$  such that  $w(\cdot) \leq 1$ ,  $w(p) + w(q) \leq 1$ .

Then gives A1–A4 and if

$$\text{A6. } H(p \cup q) = \frac{w(p) H(p) + w(q) H(q)}{w(p) + w(q)}$$

i.e.  $H(\cdot)$  satisfies the arithmetic mean–value property, then  $H(p)$  is the Shannon entropy. Higher order entropies are obtained by requiring other mean–value properties; for instance,

**Theorem 3.** (Renyi (1961)):

Let  $\varphi(x)$  be a strictly monotonic and continuous function with  $x = \varphi^{-1}(\cdot)$  denoting its inverse. Given A1–A4 and if

$$\text{A7. } H(p \cup q) = \varphi^{-1} \left[ \frac{w(p) \varphi(H(p)) + w(q) \varphi(H(q))}{w(p) + w(q)} \right]$$

we have

$H(p) = H_\gamma(p)$ , defined in (2), when  $\varphi(x) = 2^{(\gamma-1)x}$ , an exponential function,

and,

$H(p) = H_1(p)$ , the Shannon entropy, if  $\varphi(x) = ax + b$ ,  $a \neq 0$ .

Of course, when  $\varphi(x)$  is linear  $A_7$  and  $A_6$  are identical

**Cross entropy:**

Let  $H(X \cup Y)$  denote the Shannon entropy of the joint distribution of possibly dependent random variables  $X$  and  $Y$ , with marginal entropies denoted by  $H(X)$  and  $H(Y)$ . It may be verified that,

$H(X \cup Y) = H(X) + H(Y) - H(X \cap Y)$ , where

$$H(X \cap Y) = \int f(x,y) \ln[f(x,y)/f(x)f(y)] dx dy \quad (3)$$

where  $f(\cdot)$  stands for p.d.f. The expression in (3) is known as the **cross entropy (CE)** and is a very versatile measure of **dependence or association** (linear or not) for continuous, categorical and/or ordinal categorical variables. See Joe Harry (1989) for a recent account in which the superiority of CE over such measures as  $R^2$  is made abundantly clear. As we shall see in section 4 a test of independence may be based on estimated values of cross entropy. This is so since CE is a measure of "divergence" between two densities (hypotheses), a concept we now turn to.

**2.3. Distance and divergence between Distributions**

Statistical inference is best seen as a problem of measuring and judging affinities and dissimilarities among distributions. An axiomatic approach has also developed for the characterization of "ideal" measures of divergence.

Let  $p$  be a "prior" (or given) distribution and  $q$  be the conditional distribution given an occurrence of event  $E$ . What is the "information gain", or the "distance" between  $p$  and  $q$ ? Let us denote a measure of this "information gain" by  $I(q,p)$ . Similar axiom systems as were discussed earlier can be used to characterize different choices of  $I(\cdot)$ . This is a key development since metrics define all the fundamental criteria of science that are used to discuss rationality, optimality, penalty, utility, and formal

spaces. The axiomatic formulation is a powerful vehicle for analyzing the explicit as well as implicit biases that underly all such criteria. In this way a formal discussion of "consensus measures" can be conducted. To avoid repetition and in order to indicate the influence that these abstract developments have had in other disciplines, an account of these axiomatic derivations is given in later sections of this survey on economics. Here we may proceed with some examples of divergence measures as follows:

$$(i) \quad I_1(q, p) = \sum_{i=1}^n q_i \ln \frac{q_i}{p_i} = -E_q[\ln(1/q_i) - \ln(1/p_i)] \quad (4)$$

where  $p_i \neq 0$ ,  $I_1(\cdot) \geq 0$  by Jensen's inequality, and  $I_1(q, p) = 0$  iff  $q = p$ . Note that this is a **directional** (asymmetric) measure of divergence; see White (1992) for a rigorous definition for probability measures in continuous spaces.

$$(ii) \quad I_1(p, q) = I_1'(q, p) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i} \quad (5)$$

with the same properties as  $I(q, p)$  but  $q_i \neq 0$  is now required.

(iii) **Kullback - Leibler - Jeffreys measure**

$$J(q, p) = I_1'(q, p) + I_1(q, p), \text{ a symmetric measure} \\ = \sum_{i=1}^n (q_i - p_i) (\ln q_i - \ln p_i) \quad (6)$$

(iv) **The generalized k-class measure**

$$I_k(q, p) = \frac{1}{k-1} \left( \sum_{i=1}^n \frac{q_i^k}{p_i^{k-1}} - 1 \right), \quad k \neq 1 \quad (7)$$

$\lim_{k \rightarrow 1} I_k(\cdot) = I_1(q, p)$ , and

(v) **The Generalized Entropy Family:**

$$I_\gamma(q, p) = \frac{1}{\gamma(\gamma+1)} \sum_{i=1}^n q_i \left[ \left( \frac{q_i}{p_i} \right)^\gamma - 1 \right], \quad \gamma \neq 0, -1 \quad (8)$$

with  $I_0$  and  $I_{-1}$  as the limiting forms that were defined, respectively in (i) and (ii) above.

When  $X$  is a continuous random variable with  $f(x)$  as its probability density at the point  $x$ , we first note that Renyi's (1961) class of entropies can be written as

$$H_\gamma(f) = (\gamma - 1)^{-1} \log E(f(x))^{\gamma-1}, \quad \gamma \neq 1 \quad (9)$$

When  $\gamma = 1$ ,  $H_\gamma(f)$  tends to Shannon's entropy,  $-E \log(f(x))$ . Also a  $k$ -class of entropies given by Havrda and Charvat (1967) is:

$$H_k(f) = (k-1)^{-1} [1 - E f^{k-1}], \quad k \neq 1 \tag{10}$$

$= -E \log f$  at  $k = 1$ .

Note that while  $k$ -class measures in (10) satisfy the axioms A1 to A5, the entropies of Renyi satisfy A1 to A4.

Based on these entropies the measures of divergence for a continuous random variable  $X$  having densities  $f_1(x)$  and  $f_2(x)$ , correspond to (iv) and (v) as follows:

$$(iv)' \quad I_k(f_2, f_1) = \frac{1}{k-1} \int f_1^k [f_2^{1-k} - 1] \quad k \neq 1 \tag{11}$$

$$(v)' \quad I_\gamma(f_2, f_1) = \frac{1}{\gamma(\gamma+1)} \int f_2 [(f_2/f_1)^\gamma - 1] dx, \quad \gamma \neq 0 \text{ or } -1 \tag{12}$$

with  $I_1$  in (iv)' and  $I_0$  in (v)' being the continuous versions of  $I_1$  in (i), that is

$$I_1(f_2, f_1) = - \int f_2 \log(f_2/f_1) = -E_f [\log(1/f_2) - \log(1/f_1)] \tag{13}$$

We note here that for a divergence measure to be a proper distance (metric) it should be positive, symmetric and satisfy the triangular inequality. Thus the asymmetric (directed) and symmetric divergence measures given above are not proper distance measures between  $p$  and  $q$  ( $f_1$  and  $f_2$ ). This is because the divergence measure in (iii) does not satisfy the triangular inequality. But it may be verified that when  $k = 1/2$  in (iv)' we obtain a symmetric and proper distance measure as follows:

$$I_{1/2} = I_{1/2}(f_2, f_1) + I_{1/2}(f_1, f_2) = 4 [1 - \int (f_1 f_2)^{1/2} dx] = 2 \int (f_1^{1/2} - f_2^{1/2})^2 dx \tag{14}$$

Further

$$I_{1/2} = 4B(f_2, f_1) = 2M(f_2, f_1) \tag{15}$$

where

$$B(f_2, f_1) = 1 - \rho^* ; \quad \rho^* = \int (f_1 f_2)^{1/2}, \tag{16}$$

and

$$M(f_2, f_1) = \int (f_1^{1/2} - f_2^{1/2})^2 dx \tag{17}$$

are the Bhattacharya (1943), and Matuszewska (1951, 1967) distances, respectively, satisfying properties of a metric.  $\rho^*$  may serve as a measure of association since

$$0 \leq \rho^* = \int f_1^{1/2} f_2^{1/2} \leq 1.$$

Statistical properties of  $M$  are given in Kirmani (1979).

An alternative distance measure is a Hilbert  $L_2$ -norm distance and has been used in the nonparametric hypothesis testing literature

$$L_2 = H(f_2, f_1) = \int [(f_2 - f_1)^2]^{1/2} \quad (18)$$

This is an example of  $L_p$ -norm distance,  $L_p = \{ \int |f_2 - f_1|^p \}^{1/p}$ . The important case  $p = 1$  corresponds to the well known Kolmogorov's (1963) distance.

Several of these distance measures are used in cluster analysis. For an application and further discussion see section 8 below. In section 6 we will discuss further characterization theorems (axiom systems) for some of these measures in the terminology of modern welfare theory. Aczel and Daroczy (1975) is a good source for relevant derivations and technical details.

### 3. Maximum Entropy, MLE, Min $\chi^2$ , and LS

All "rational" decisions and measurement procedures in science are justified or formalized on the basis of optimization relative to suitable criteria and restrictions. But what axiom systems justify any particular criterion? Are there particular criteria that are more easily (universally) justifiable than others? The axiom systems described earlier in this paper are meant to be suggestive of answers to these questions. Generally, Kullback - Leibler measure is the criterion of choice if rather plausible branching/aggregation consistency restrictions are imposed. We will focus on this measure of divergence in much of this paper, but see section 6 for generalizations.

#### 3.1 The ME principle

Maximum Likelihood (ML) method begins with the specification of a distribution function (family  $f(y, \theta)$  say) and then relies upon data to infer a particular member ( $\theta$ ). If  $f(\cdot)$  is given we know all of its moments. But if we are given the values of some

moments (or moment functions), we have less information as there are generally many  $f(\cdot)$  compatible with such moments. The Maximum Entropy (ME) and Minimum Information (MI) principles state that we should find a distribution in such a way as to minimize the (inadvertant) use of information other than is explicitly available. We may regard this as a definition of **disciplined inference**. Almost all statistically oriented problems can be translated into this type of "inverse problem". It is formally as follows:

Let  $f_0(\cdot)$  be the prior density, if any, and solve the constrained optimization:

$$\text{Min } -\int f(x) \ln \{f(x)/f_0(x)\} dx \quad (19)$$

subject to

$$\int f(x) dx = 1, \text{ and } E g_r(x) = a_r, \quad r = 1, 2, \dots, m$$

Some popular  $g(\cdot)$  functions are  $x$ ,  $x^r$ ,  $(\ln x)^r$ ,  $\ln(1+x^r)$ ,  $|x - Ex|^r$ ,  $\sin x$ ,  $\cos x$ , etc. The general solution, known as the ME distribution, is:

$$f(x) = f_0(x) \exp [-\lambda_0 - \lambda_1 g_1(x) - \dots - \lambda_m g_m(x)] \quad (20)$$

where  $\lambda_j$  are obtained so as to satisfy the constraints. The unknown moments  $a_r$  must be estimated in practice, either by ML or the method of moments. This is the sense in which one may interpret ME based techniques as "disciplined" method of moments. When such estimates are substituted for  $a_r$  the corresponding ME density is denoted by  $\hat{f}(x)$ . Also, there are many numerical algorithms for implementing the above "inverse problem". One is demonstrated in Zellner and Highfield (1988), but this algorithm did not work in most of the cases we tried. In Maasoumi and Zhang (1992b) other algorithms are proposed, including one for the computation of **generalized entropy**. We are computing new measures of "volatility" in financial returns data based on these entropies. See section 4 below for further suggestions. When  $f_0$  is the uniform density the ME distribution simply maximizes the Shannon entropy. Tables 1-2 provide some examples of well known distributions derived as ME distributions under different constraints.

Table 1  
 MAXIMUM-ENTROPY DISCRETE-VARIATE PROBABILITY  
 DISTRIBUTIONS

Range of Variate	Specified Moments	Prior Dista $q_i$	ME-MI Dista $P_i$	Name
1,2,...,n	-	-	$\frac{1}{n}$	uniform
1,2,...,m	mean m	uniform	$ap^i$	geometric
0,1,2,...,n	mean	$\binom{n}{i}$	$\binom{n}{i} p^i q^{n-i}$	binomial
0,1,2,3,...	mean m	$(i!)^{-1}$	$\frac{e^{-m} m^i}{i!}$	Poisson
1,2,3,...	mean m	$i^{-1}$	$\frac{1}{m(1-q)} \frac{q^i}{i}$	Log Series
1,2,3,...	mean m	$i^{-d}$	$\frac{\sum_{i=1}^{\infty} q^i}{i^d}$	general geometric

### 3.2. Relation to MLE and $\text{Min } \chi^2$

Let us consider the discrete case for simplicity and direct connection to potential applications in discrete variable models. Let  $P=(p_1, p_2, \dots, p_n)$  be an unknown distribution for  $x$ , and  $P_0=(p_{10}, p_{20}, \dots, p_{n0})$  its ME counterpart. Let  $S$  and  $S_m$  be their respective entropies, and  $\Delta S = S_m - S$ . Jaynes (1979) provided the following "concentration theorem":

Theorem: In  $N$  random trials,  $2N\Delta S$  is asymptotically distributed as chi-squared with  $k = n-m-1$  degrees of freedom.

This theorem allows confidence intervals to be constructed for observed (estimated, hypothesized) distributions; It can be shown that:

Table 2

MAXIMUM-ENTROPY CONTINUOUS-VARIATE PROBABILITY DISTRIBUTIONS

a) <u>Range</u> $(-\infty, \infty)$	<u>Distribution</u>
<u>Specified Moments</u>	Does not exist
$E(x)$	
$E(x) = m, E(x - m)^2 = \sigma^2$	$N(m, \sigma^2)$
$E(x - \bar{x})^2 = \sigma^2$	$N(m, \sigma^2)$ (m arbitrary)
$E(x^r) = a_r$	Does not exist if k is odd
$r = 1, 2, \dots, k$	$f(x) = \exp[-\lambda_0 - \lambda_1 x - \dots - \lambda_k x^k]$ , if k is even
$E( x ) = \sigma$	Laplace
$E( x - m ) = \sigma$	Laplace with mean m
$E(x) = m, E( x - m ) = \sigma$	Laplace with mean m
$E \ln(1 + x^2)$	Generalized Cauchy
b) <u>Range</u> $[0, \infty)$	
$E(x)$	exponential
$E(x), E(\ln x)$	gamma
$E(x), E(\ln(1+x))$	beta
$E(\ln x), E(\ln x)^2$	log normal
$E \ln(1+x^2)$	unilateral generalized Cauchy
$E(x) = m, E(x^2) = \sigma^2$	Truncated normal if $\sigma^2 < m^2$ exponential if $\sigma^2 = m^2$ does not exist if $\sigma^2 > m^2$
c) <u>Range</u> $[0, 1]$	<u>Distribution</u>
<u>Specified Moments</u>	uniform
none	truncated exponential
mean	truncated normal or truncated u or uniform depending on prescribed values
$E(x), E(x^2)$	beta
$E(\ln x), E(\ln(1-x))$	

$$\begin{aligned} \Delta S &= -\sum p_{i0} \ln p_{i0} + \sum p_i \ln p_i = \sum p_i \log(p_i/p_{i0}) \\ &= \frac{1}{2} \sum (p_{i0} - p_i)^2 / p_{i0} + \dots = \Delta I \end{aligned} \quad (21)$$

$$\begin{aligned} N \Delta I &= N |I - I_{\min}| = \frac{1}{2} \sum (N p_{i0} - N p_i)^2 / N p_{i0} + \dots \\ I &= \sum p_i \log(p_i/q_i); \quad I_{\min} = \sum p_{i0} \log(p_{i0}/q_i) \end{aligned} \quad (22)$$

with q representing the "prior" distribution, if any.



Interpreting  $Np_i = x_i$  as the observed frequencies and  $Np_{i0}$  as the expected (predicted or hypothesized) frequencies, the above expressions form the basis for the proof of Jaynes' theorem, and also the following equivalence result for MLE:

Let  $P(\cdot; \mu)$  be as defined above and  $x_i, i = 1, 2, \dots, n$ , be the observed frequencies in  $N$  independent trials. The log likelihood function is

$$\begin{aligned} \ln L(\mu; x_1, x_2, \dots, x_n) &= \ln(N! / x_1! x_2! \dots x_n!) p_1^{x_1} p_2^{x_2} \dots p_n^{x_n} \\ &= [\ln(\dots) + \sum x_i \ln(x_i/N)] - \sum x_i \ln(x_i/Np_i) \\ &= \ln C - \sum x_i \ln(x_i/Np_i) \end{aligned} \quad (23)$$

By Shannon's inequality the second term is non-negative and vanishes iff  $p_i = x_i/N$ . Thus  $C = L_{\max}$  over all  $p_i$ . Again, expanding the second term of  $\ln L(\cdot)$  above, we have:

$$\begin{aligned} \Delta \ln L &= \ln L_{\max} - \ln L = \frac{1}{2} \chi_k^2 + \dots, \text{ where,} \\ \chi_k^2 &= \sum (Np_i - x_i)^2 / x_i \approx \sum (Np_i - x_i)^2 / Np_i \approx \sum (Np_i - x_i)^2 / Np_{i0} \end{aligned}$$

We conclude that, asymptotically:

$$2N\Delta S = 2N\Delta I = 2\Delta \ln L = \chi_k^2 \quad (24)$$

This is the essential connection between Min chi-square and ML criterion on the one hand, and axiomatically justified information criteria of "divergence" between distributions, models and hypotheses, on the other. There is also a related interpretation of MLE as an information theoretic estimation procedure, as follows:

Let  $f(y)$  and  $g(y, \theta)$  be two probability densities for the random variable  $y$ , with unknown parameter  $\theta$ , and the Kullback-Leibler divergence given by:

$$\begin{aligned} I(f;g) &= \int f \ln(f/g) dy \\ &= \int f(y) \ln f(y) dy - \int f(y) \ln g(y) dy \end{aligned} \quad (25)$$

If we are given  $f(y)$ , for instance when it is observed, and if  $g(\cdot)$  is known upto  $\theta$ , to minimize  $I(\cdot)$  would be equivalent to minimizing the last term (cross-entropy).

Arranging the data in ascending order, we have

$$f(y) = 1/n, \text{ for } y_i \leq y \leq y_{i+1}, i = 1, 2, \dots, n.$$

and if we minimize the empirical cross-entropy,  $-\frac{1}{n} \sum \ln g(y_i, \theta)$ , w.r.t.  $\theta$ , we would

be maximizing  $\frac{1}{n} \log L(y_1, y_2, \dots, y_n; \theta)$ , the log likelihood of  $n$  independent observations. Thus

$$\text{MLE}(\theta) \equiv \text{Min Info. Divergence (MID) estimator.}$$

Edgeworth-type approximations to entropy functionals are discussed in Maasoumi and Theil (1979). Note that information criteria are not dependent on dimension and/or parametric forms, making them especially suitable in nonparametric settings and for non-nested hypothesis testing; see I.J. Good (1963) for an early discussion.

### 3.3 Financial Entropy.

Recently Stutzer (1992) has given an interesting interpretation to the minimized Kullback-Leibler divergence between a risk neutral probability measure,  $d\nu$ , and the actual measure  $d\mu$  of the relative returns on assets in an arbitrage-free market. This "financial entropy" is a useful indicator of the information gained from the observed returns data regarding the degree of risk adjustment necessitated by an absence of arbitrage and the existence of risk premia. Stutzer relies upon a relative ME inverse problem for  $N$  assets with the constraints:

$$E_{\nu}[X_i] = E_{\mu} \left\{ X_i \frac{d\nu}{d\mu} \right\} = 0, \quad i = 1, \dots, N$$

$X = (R/r) - 1$ ,  $R$  is the gross real return on an asset, and  $r$  is the gross real interest rate. When  $\mu$  is uniform the minimization of this relative entropy (Kullback-Leibler divergence) is equivalent to the maximization of Shannon's entropy with respect to  $d\nu$ , the quantity measuring the uncertainty in the measure  $\nu$ . See Shore and Johnson (1980) for a recent axiomatic justification of the Kullback-Leibler measure of relative entropy. See Stutzer (1992) for further references to applications in finance theory.

### 3.4 Pseudo-true models and Parameters.

An increasing number of econometricians have begun to accept by their deeds that non-experimental data are inevitably analyzed by misspecified models. Models may be misspecified theoretically, statistically, or both. If models are misspecified in an indeterminate manner, then we should not be aiming at the discovery of "true data generating processes". As has been argued, for example see Maasoumi (1988a), it is

useful and perhaps inevitable to define and work with statistically accessible populations about which reasonable inferences may be drawn. Information theory and the ME principle afford us a means of defining such objects of inference in a way that combines the sample information as well as a priori model classes from which a member is then selected. It is often the case that such pseudo-realities are much more successfully accessed by statistical inference and remove a common misunderstanding about learning about scientific models rather than statistically unconfirmable phenomena. A simple example may be worthwhile:

Let  $Y$  be the variable(s) we wish to learn about. Let  $F(Y|X)$  represent a conditional statistical model class which need not be parametric or specified to any degree. A priori there is usually no compelling reason for parametric specifications of  $F(Y|X)$ , but it is often scientifically convenient to think of another distribution,  $G(Y|X) = N(X\beta, \sigma^2\Omega)$ , say, as a formal approximation. Note, however, that attribution of meanings to such parameters as  $\beta$  at this point bears no necessary relationship to  $F(\cdot)$  and is not required. Indeed  $G(\cdot)$  may be allowed to be non-parametric itself. Information theory suggests a first step in which we may first select an appropriate member of the family presented by  $G(\cdot)$ , i.e., a choice of  $\beta$  and  $\sigma^2$  in the parametric case, which is in some sense the "closest" to  $F(\cdot)$  before any estimation is to take place. For instance, if the kullback-Leibler-Jeffreys' measure between  $F$  and  $G$  is minimized with respect to  $\beta$  and  $\sigma^2$ , we obtain:

$$\beta_0 = (X'X)^{-1}X'\mu_T, \text{ and } \sigma_0^2 = \frac{1}{T} \mu_T' M_X \mu_T + \frac{1}{T} \text{tr } \Omega \quad (26)$$

where  $T$  is the dimension of  $Y$ ,  $\mu_T = E_T(Y|X)$ ,  $M_X$  is the well known "fundamental regression matrix", and  $(\beta_0, \sigma_0^2)$  are known as the "pseudo-true" parameters; see Sawa (1978). The corresponding member of  $G(Y|X)$  class,  $G_0(Y|X)$ , is what we refer to as the statistically accessible model. A simple moments estimate of the mean of  $Y$  may replace the unknown  $\mu_T$  to provide usually consistent estimators of  $\beta_0$  and  $\sigma_0^2$  (i.e.,  $G_0$ ). For instance, OLS estimates in the linear regression model are seen to be more commonly consistent for the pseudo-true parameters whatever  $\beta$  and  $\sigma^2$  may be.

Indeed, it is not clear why the latter parameters should be of interest! Learning about  $\beta$  and  $\sigma^2$  is not a well-posed statistical question. See section 5 for the related work of Rissanen which advocates the encoding of both the data and model classes from which an ideal member is then chosen. Like  $G_0$  this member is a data-dependent scientific target of inference.

The definition of pseudo-true parameters may be given in more general "approximate" or "misspecified" models than the linear one in  $G(\cdot)$  above. The following is from H. White (1992) where appropriate conditions are fully laid out:

Let  $g(Y_t, \theta)$  be the approximate probability model and  $f(Y_t)$  the true one. The parameter  $\theta^*$  which minimizes the Kullback-Leibler divergence between these two densities is the pseudo-true and  $g(Y_t, \theta^*)$  is the pseudo-true model. It is also true that  $\theta^*$  is the solution to

$$\text{Max}_{\theta \in \Theta} E[\log g(Y_t, \theta)] \quad \Theta = \text{the admissible values of } \theta.$$

And a commonly consistent and asymptotically normal estimator of  $\theta^*$  is obtained from maximization of the mean quasi-log likelihood function for a sample of T observations

$$\text{Max}_{\theta \in \Theta} T^{-1} \sum_t \log g(Y_t, \theta)$$

**4. ME regressions, estimation, and testing.**

**4.1 Estimation of Econometric Functions Using ME density Functions**

In econometrics we study various functions of interest (conditional moments), for example regression functions, heteroskedasticity functions, autocorrelation functions. Usually, these functions are studied by assuming parametric forms. But one can avoid *a priori* specifications by using the ME density defined above. This motivation has been suggested and developed in Maasoumi (1979, 1986b and 1988b). Ryu (1991) has further developed this idea for conditional moments as regression functions. We note that an unknown function can be transformed so as to satisfy the properties of a density function. Then the ME solution is obtained which in turn obtains the optimal functional form. The following example is given by Ryu (1991):

For the two variables  $Y$  and  $X$  we note that the regression function is

$$m(x) = E(Y|X=x) = \int yf(y|x)dy = \frac{\int y f(y, x) dy}{f(x)} \quad (27)$$

Similarly, for the variance (conditional heteroskedasticity) and higher moments:

$$\sigma^2(x) = V(Y|X=x) = E(Y^2|X=x) - m^2(x) \quad (28)$$

$$= \frac{\int y^2 f(y, x) dy}{f(x)} - m^2(x) \quad (29)$$

The numerator of the regression function  $m(x)$  can be written as

$$h(x) = \int_y yf(y,x)dy = m(x)f(x) \quad (30)$$

Note that, since  $\int h(x)dx = \int yf(y,x)dy = EY$ ,  $h^*(x) = h(x)/EY$  is a density function.

Thus

$$\max_{h^*} [-\int h^*(x)\log h^*(x)dx] \equiv \max_h [-\int h(x)\log h(x)dx]$$

and we can derive the ME estimate of  $h^*$  under such restrictions  $\int x^r h^*(x)dx = a_r$ ,  $r=1, \dots, m$ . We may obtain an expression for  $h^*(x)$  as in section 3.1, with  $g_r(x) = x^r$ , and an estimator of  $a_r$  such as  $\hat{a}_r = n^{-1} \sum_{i=1}^n x_i^r / \bar{y}$ , where  $x_i$  and  $y_i$  denote the sample observations on  $X$  and  $Y$ . Further,  $\hat{h}(x) = (\hat{E}y) \hat{h}^*(x)$  with  $\hat{E}y$  replaced by an estimate such as  $\bar{y}$ . Finally the ME estimator of the regression function  $m(x)$  is

$$\hat{m}(x) = \frac{\hat{h}(x)}{\hat{f}(x)} \quad (31)$$

where  $\hat{f}(x)$  is the ME density of  $x$  derived as described in section 3.1.

Maasoumi (1985, 86b) proposed the use of Generalized Entropy (GE) method for generating unknown economic and regression functions without side restrictions. "Ideal" aggregation functions, such as generalized geometric and hyperbolic functions, are derived in this spirit (see section 7). Ryu (1991) has given some examples of ME regression functions based on the Kullback-Leibler measure (rather than GE) but with moment restrictions. Table 3 is based on his results.

Table 3

ME functions with known restrictions

Functional forms	$g_1(x_1), g_{ij}(x_1, x_2), g_{m_1}(x_1)g_{m_2}(x_2)$
<b>(a) Exponential polynomial</b> $y(x_1, x_2) = \exp\left[\sum_{n_1, n_2=0}^{N_1, N_2} a_{n_1, n_2} x_1^{n_1} x_2^{n_2}\right]$	For $m_1 = 0, \dots, N_1$ and $m_2 = 0, \dots, N_2$ $g_{m_1}(x_1)g_{m_2}(x_2) = x_1^{m_1} x_2^{m_2}$
<b>(b) Cobb-Douglas function</b> $\log y(x_1, x_2) = a_0 + \sum_{i=1}^2 a_i \log x_i$	For $i = 1, 2$ $g_0(x) = 1, g_i(x_i) = \log x_i$
<b>(c) Translog function</b> $\log y(x_1, x_2) = a_0 + \sum_{i=1}^2 a_i \log x_i + \sum_{n_1, n_2=1}^2 a_{n_1, n_2} (\log x_{n_1})(\log x_{n_2})$	For $i = 1, 2; m_1 = 1, 2$ , and $m_2 = 1, 2$ $g_0(x) = 1, g_i(x_i) = \log x_i$ , and $g_{m_1}(x_1)g_{m_2}(x_2) = (\log x_{m_1})(\log x_{m_2})$
<b>(d) Generalized Cobb-Douglas</b> $\log y(x_1, x_2) = a_0 + \sum_{i,j=1}^2 a_{ij} \log(x_i + x_j)/2$	For $i, j = 1, 2$ $g_0 = 1, g_{ij}(x_i, x_j) = \log(x_i + x_j)/2$
<b>(e) Generalized Leontief</b> $z = a_0 + \sum_{i=1}^2 a_i \sqrt{x_i} + \sum_{i,j=1}^2 a_{ij} \sqrt{x_i x_j}$	For $i, j = 1, 2$ $z = \log y(x_1, x_2), g_0 = 1, g_i = \sqrt{x_i}$ , $g_{ij}(x_i, x_j) = \sqrt{x_i x_j}$
<b>(f) Fourier flexible form</b> $z = a_0 + \sum_{i=1}^2 a_i x_i + \sum_{i,j=1}^2 a_{ij} x_i x_j + \sum_{n_1, n_2=0, n_1+n_2 \leq 2}^{N_1, N_2} b_{n_1, n_2} x_1^{n_1} x_2^{n_2} \exp[i\pi x_1] \exp[i\pi x_2]$	For $i, j = 1, 2$ and $(m_1, m_2) = (1, 0), (0, 1), \dots, (N_1, N_2)$ $z = \log y(x_1, x_2), g_0(x) = 1$ $g_i(x_i) = x_i, g_{ij}(x_i, x_j) = x_i x_j$ , and, $g_{m_1}(x_1)g_{m_2}(x_2) = \exp[i\pi x_1] \exp[i\pi x_2]$
<b>(g) Minflex-Laurent translog form</b> $\log y(x_1, x_2) = a_0 + 2 \sum_{i=1}^2 a_i \log x_i + \sum_{i=1}^2 a_{ii} (\log x_i)^2 - \sum_{(i,j) \in S} b_{ij} (\log x_i)(\log x_j) + \sum_{(i,j) \in S} c_{ij} (\log x_i)^{-1} (\log x_j)^{-1}$	For $i = 1, 2$ and $S = \{(i, j) : i \neq j, i, j = 1, 2\}$ $g_0 = 1, g_i(x_i) = 2 \log x_i, g_{ii}(x_i, x_i) = (\log x_i)^2$ , for $i = 1, 2$ , $g_{ij}^b = (\log x_i)(\log x_j), g_{ij}^c = (\log x_i)^{-1} (\log x_j)^{-1}$ for $(i, j) \in S$

The ME estimator of higher order conditional moments of  $Y$  can be determined similarly. For example, the ME estimator of the variance is

$$\hat{\sigma}^2(x) = \frac{\hat{\psi}''(x)}{\hat{f}(x)} - \hat{m}^2(x) \quad (32)$$

where  $\hat{\psi}(x)$  is the density obtained from

$$\max_{\psi} \left[ - \int \psi(x) \log \psi(x) dx \right]$$

satisfying  $E x^r = a_r$ . Note that  $\psi(x) = \int y^2 f(y, x) dy / E y^2$  and an estimate of  $a_r$  such as  $n^{-1} \sum_{i=1}^n x_i^r y_i^2 / (n^{-1} \sum y_i^2)$  must be used in practice. The estimator  $\hat{\sigma}^2(x)$  can be used to study volatility in economics and finance. One may contrast this ME variance estimate with entropy itself as a measure of "volatility", see Maasoumi and Zhang (1992a).

The expressions for  $\hat{m}(x)$  and  $\hat{\sigma}(x)$  will depend on the choice of  $g_r(x)$  in the moment conditions  $E g_r(x) = a_r$ . In the above discussion  $g_r(x) = x^r$ . As in Ryu (1991), however, if  $g_r(x) = P_r(x)$ , some polynomial in  $x$ , we obtain the "orthonormal basis regression estimators" of  $m(x)$ . Further, if we use  $g_r(x) = \exp(irx)$ , a trigonometric function,  $\hat{m}(x)$  will be Gallant's (1981) Fourier flexible form. See table 3 above for other examples and Ryu (1991) where some asymptotic properties of orthonormal basis estimator is given. The choice of the "smoothing function"  $g(x)$  and the "smoothing parameter"  $m$  (the number of restrictions) poses questions that are akin to those in kernel density estimation. A large value for  $m$  is equivalent to using a great deal of a priori information and can lead to the so-called "over-fitting" problem in practice. Our ability to obtain reliable sample estimates of higher moments of random variables should, however, impose natural limits in practice which are lacking in alternative non-parametric approaches.

#### 4.2 Semi-Parametric Models, Adaptive and Robust Estimation

In addition to the nonparametric specification of the regression type functions by using the ME density estimator, we can also use the ME technique for robust estimation of semiparametric models. For example, if  $y_i = x_i \beta + u_i$  where  $u_i$  is an i.i.d

random error with an unknown density  $f(u)$ , then the maximum likelihood estimation of  $\beta$  can be carried out by using the ME density  $f(u)$ . To see this, write the (log) likelihood function as

$$\log L = \sum_{i=1}^n \log f(y_i - x_i \beta) \quad (33)$$

and

$$\frac{\partial \log L}{\partial \beta} = S(\beta) = 0 = \sum_{i=1}^n \psi(u_i) x_i \quad (34)$$

where  $\psi(u_i) = -f'(u_i)/f(u_i) = -\partial \log f(u_i)/\partial u_i$  is the score of  $f(u_i)$ .  $f'/f$  plays an ubiquitous role in estimation of location in modern statistics, as well as in score tests, see Joiner and Hall (1983), Bera and Ng (1991), Manski (1984) and Koenker (1982). For instance, the two-step estimator of  $\beta$  can be obtained as

$$\hat{\beta}^* = \hat{\beta} + (I(\hat{\beta}))^{-1} S(\hat{\beta}) \quad (35)$$

where  $I(\hat{\beta})$  is the Fisher information matrix evaluated at an initial consistent estimator  $\hat{\beta}$ . Note that  $I(\hat{\beta})$  and  $S(\hat{\beta}) = \sum_{i=1}^n \psi(u_i) x_i$  can be obtained by using the ME density  $\hat{f}(u_i)$  and its derivative in  $\psi(u_i)$ , where  $u_i = y_i - x_i \hat{\beta}$ .

Another semiparametric model is  $y_i = x_i \beta + u_i$  where  $V(u_i | x_i) = \sigma^2(x_i)$  is unknown. A two step generalized least squares estimator of  $\beta$  can be obtained by using the ME estimator of  $\sigma^2(x_i)$  in the first stage. Such estimators may be contrasted with popular heteroskedasticity and/or autocorrelation consistent covariance estimators proposed by H. White, Newey and West, D. Andrews and others. Theil and others, for example see Theil and Laitinen (1980) and Theil and Fiebig (1984), have studied the computation of ME moments and densities extensively. From this work and also work done by Scof (1991) and others, we also learn that ME second moments may be successfully used to deal with problems of multicollinearity and "undersized samples" in which the usual sample second moments are singular.

#### 4.3 Tests using information criteria

There is a long but less than fully developed tradition of using entropy to formulate hypotheses. The essential idea is to estimate the difference between the entropies

obtained under different hypotheses and models. The distribution of a suitably standardized function of this entropy divergence would provide the basis for inferences. For example, recently Robinson (1991) derived the normal limiting distribution of the Kullback–Leibler criterion and showed its consistency in one-sided tests of nested hypotheses. He considered tests of independence and the random walk hypotheses. A difficult form of weighting (trimming) is necessary to establish Robinson's asymptotic distributions which rely on kernel estimation of unknown densities. We propose ME densities in place of the latter, and bootstrap methods for standardizing transformations and inference in small sample situations. Sin and White (1992) have proved the same result as Robinson but do not require the weighting scheme employed by the latter. Some examples are as follows:

(i) **Test for normality:**

Whenever the entropy of a parametrically specified model is known, its estimate may be contrasted with a nonparametric estimate of the entropy of the sample. Suppose we wish to test the following hypothesis:

$$\begin{aligned} H_0 : f(x) &= f_0(x) = N(\mu, \sigma^2) \\ H_1 : f(x) &\neq f_0(x). \end{aligned}$$

This problem can arise in testing for the normality of an economic variable or the error term in regressions. To test this hypothesis we can consider a test statistic which is based on the entropy difference or the divergence measure  $I_1$  given in section 3.1. This is

$$d(f) = H_0(f) - H(f) = I_1(f, f_0(x)) \quad (36)$$

where  $H_0(f)$  is a parametric evaluation of the entropy of  $f$  under  $H_0$ , given by

$$H_0(f) = - \int f_0(x) \log f_0(x) dx = \log \sigma + \frac{1}{2} + \log \sqrt{2\pi} \quad (37)$$

while  $H(f)$  is a non-parametric evaluation of  $H(f) = -E \log f(x)$ , the entropy of the alternative. In practice we replace  $H_0(f)$  by  $\hat{H}_0(f) = \log \hat{\sigma} + \frac{1}{2} + \log \sqrt{2\pi}$  where  $\hat{\sigma}^2 = \sum (x_i - \bar{x})^2 / (n-1)$ , and  $\hat{H}(f) = - \sum_{i=1}^n \log \hat{f}(x_i) / n$  where  $\hat{f}$  is the usual maximum

entropy estimator. One may consider kernel density estimators of  $f$  in order to calculate  $\hat{H}(f)$ . Vasicek (1976), however, introduced the following "empiric entropy" estimator:

$$\hat{H}(f) = -\frac{1}{n} \sum_{i=1}^n \log \left( \frac{n}{2m} (y_{i+m} - y_{i-m}) \right) \tag{38}$$

where, given the observations  $x_1, y_1 \leq y_2 \leq \dots \leq y_n$  denote the order statistics,  $y_i = y_i (i < 1)$  and  $y_i = y_n (i > n)$ . Using  $\hat{H}(f)$  in  $d(f)$  Vasicek studied the properties of the test statistic and provided the critical points of the test. The case  $m=1$  is known as Theil's entropy estimator. The latter should be distinguished from Theil's inequality measure for the size distribution of a variable; See section 6.1.3 below.

(ii) Test for Independence:

Suppose we wish to test the independence of two random variables  $Y$  and  $X$ . Then

$$H_0 : f(y,x) = f(y)f(x)$$

$$H_1 : f(y,x) \neq f(y)f(x)$$

A test statistic based on the entropy divergence is

$$d(H_0, H_1) = H_0(f) - H(f) \tag{39}$$

$$= -\iint f(y)f(x) [\log f(y) + \log f(x)] dy dx + \iint f(y,x) \log f(y,x) dy dx \tag{40}$$

$$= H(f(y)) + H(f(x)) - H(f(y,x)) \tag{41}$$

$$= I_1(f(y,x), f(y)f(x)), \tag{42}$$

where  $I_1$  is a measure of divergence of  $f(y)f(x)$  from  $f(y,x)$ . Thus testing for  $H_0$  implies testing for  $d = I_1 = 0$ .

In practice  $d$  can be calculated by substituting nonparametric estimators of  $H(f(y))$  and  $H(f(x))$  as given above, and of  $H(f(y,x))$  by

$$\hat{H}(f(y,x)) = -\frac{1}{n} \sum_{i=1}^n \log \hat{f}(y_i, x_i) \tag{43}$$

The asymptotic normality of  $d$  based on these estimates is established in Robinson (1991). But Robinson employs kernel density estimation to obtain the unknown densities, see also Ahmad and Lin (1976) and Sin and White (1992). It would be useful

to examine the alternative based on ME densities, as well as the small sample properties of the tests by bootstrap methods. In particular, inferences based on the normal distribution must be viewed with care.

Tests for the symmetry of a density function,  $H_0 : f(x) = f(-x)$ , vs  $H_1 : f(x) \neq f(-x)$ , and tests for equality of densities,  $H_0 : f_Y(x) = f_X(x)$  vs  $H_1 : (\text{not } H_0)$ , can also be developed by using the above procedures.

(iii) **Residual based tests:**

A test for linearity against a ME regression  $m(x)$  can be developed by an adaptation of the Ramsey's RESET or other residual based procedures. See Pagan and Hall (1983). But instead of polynomial or non-parametric regressions, suitable functions of the residuals from an ME regression and residuals from a linear regression may be contrasted. Indeed the contrast criteria themselves may be chosen from information theory although the appropriate distribution theory is not yet available. Given the accumulated evidence on the performance of asymptotic theory in small samples, however, it would seem both necessary and desirable to derive better approximations by (e.g.) bootstrap methods.

Tests for other hypotheses, such as serial correlation and heteroskedasticity, may be developed where regression errors are given the ME density. Rao's score principle can be adopted for a class of non-Gaussian distributions, see Bickel (1982), and Bera and Ng (1991). The latter paper considers several different estimators of the score function. It would seem worthwhile to consider the ME estimator of the score in these cases. In fact the most versatile and currently popular testing procedures are the Lagrangian Multiplier (Score) tests. Clearly, this entire area of inference awaits revision because of the less than satisfactory performance of the LM test in small samples. It is well known that much of the difficulty can be traced to the estimation of the score function and the information matrix. I would conjecture that the ME alternative would be more successful since ME density estimates tend to provide for better and smoother tail area estimates. I would expect this superiority to materialize

even over other non-parametric (kernel) density estimates which tend to do badly in the tails. Once again, computer intensive evaluations of small sample distributions of test statistics should guide the applications of these new test procedures.

5. Akaike and Rissanen criteria, Fisher's information matrix, higher entropies and geometries.

#### 5.1 Model selection and fit criteria

The decision on whether a model has adequately fit the available data, or whether it predicts well, or how it compares with competing models is perhaps even more important in the social sciences than in "experimental" areas. Observational data require multiple regression type "controls" as well as consideration of many competing models that are empirically and/or theoretically plausible. Such decisions require criteria of "goodness" which can be taken to mean criteria for measuring "closeness" to some ideal. But this is precisely the task for information criteria of the type described in this paper and chosen for its main theme. Many of these criteria either subsume the well known (e.g) mean squared error type functions or compete with such traditional measures. The most important difference between them is that information functions are typically dimension-independent and non-parametric if desired, making them far more effective in comparison of disparate (e.g., non-nested) models. Section 4 contained some examples of these criteria when the closely related question of hypothesis testing was addressed. Below several criteria are briefly described, but a full list would be virtually endless.

Based on the Kullback-Liebler measure, Akaike (1973) proposed the most widely analyzed information measure of model (variable) choice. This measure has been used in both time series and regression contexts. Akaike's measure must be credited with being the first widely known measure that attempts to address the divergent requirements of model complexity and estimation accuracy (fit, likelihood). But as will be seen, Akaike's measures take parsimony as an indicator of "complexity", whereas recent measures proposed by Rissanen attempt to go further in this regard.

Akaike (1973) introduced the following measure (AIC):

$$\text{AIC} = (-2)\log(\text{max Likelihood}) + 2(\text{number of independently adjusted parameters within the model}) \quad (44)$$

The model of choice would minimize this criterion. This would penalize lack of parsimony, and depend on the likelihood to measure fit and estimation accuracy. In the context of variable selection in regression the expression for AIC is:

$$\text{AIC} = -2 \ln L(b_1 | y) / T + 2K_1 / T \quad (45)$$

where  $L(\cdot)$  denotes the likelihood,  $y$  is the dependent variable,  $b_1$  is the estimate of the  $K_1$  independent (unrestricted) regression coefficients, and  $T$  denotes the sample size. Denoting the corresponding regressor matrix  $X_1$  and  $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$ , (45) may be rewritten as follows

$$\text{AIC}_{(\text{known } \sigma^2)} = y'M_1y/T + 2\sigma^2K_1/T \quad (46)$$

The relationship between this measure and several others may be noted. This is often a matter of how the unknown coefficients are estimated. For instance, if  $\sigma^2$  is estimated by its usual unbiased estimator,  $\hat{\sigma}^2 = y'My/(T-K)$  with  $X$  denoting all of the  $K$  regressors being considered and  $M$  defined similarly to  $M_1$ , AIC becomes identical to Mallows  $C_p$  criterion. On the other hand, a prediction criterion proposed by Amemiya (1980) is obtained from AIC when  $\sigma^2$  is estimated by  $y'M_1y/(T-K_1)$ .

Sawa (1978) proposed a Best Information Criterion (BIC) which is also based on the Kullback–Liebler divergence but does not require the existence of a "true" model. A pseudo-true model is first defined. The role of information criteria in defining pseudo true parameters and models is quite significant by now. See section 3 of this paper for more detail. See White (1990) and Sin and White (1992) for recent examples. For instance, the latter paper develops a penalized likelihood criterion for model selection for sufficiently regular dependent processes in models that may be nested or non-nested, linear or nonlinear, and possibly misspecified. They give sufficient conditions under which their criterion selects, with probability one or approaching one, the model that attains the lower average Kullback–Leibler divergence. They show that Akaike's, Schwarz' and Hannan–Quinn information criteria are special cases.

Akaike and others have extended AIC in both the frequentist and Bayesian settings. These criteria have become sufficiently popular in econometrics to be presented in textbooks; e.g., see Judge et al. (1985). But it would be inaccurate to think that a closure is at hand in the debate regarding the separation of hypothesis testing which Akaike and many others consider as riddled with subjective choices, on the one hand, and those who do not think there is a significant new element in these information criteria. See Leamer (1979).

A very promising and seemingly general criterion for the measurement of stochastic complexity of models has been proposed by Rissanen. See Rissanen (1987, 1988, 1989). Rissanen's work builds upon notions of algorithmic complexity developed by Solomonoff, Kolmogorov and Chaitin. In order to employ the Minimum Description Length (MDL) criterion developed for encoded phenomena, data and the model class are first encoded. Rissanen (1989) provides a review of several techniques for encoding that reveals once again a certain vague universality for  $-\ln(\text{probability})$  as a store of information, a "prefix" code. Once this coding is achieved, codes and code length rather than models become the primary objective of choice. One then chooses from within a model class by minimizing the code length. Note that, like the AIC, both the data and model characteristics are taken into account. A slightly simpler and more limited concept of "complexity" is, of course, present in AIC. Rissanen's criterion is applicable to more complicated models. He has provided many examples in his work, including the regression problem discussed above.

Let us see the connection between MDL and stochastic complexity for the probabilistic model class

$$M_k = [P(y|\theta), \pi(\theta)] \quad (47)$$

where  $P(\cdot)$  is the probability function (or density),  $\pi$  is the prior on  $\theta$  which may or may not exist and is not relevant asymptotically, and  $\theta$  is a  $k$ -component parameter vector. For a sample of  $n$  observations on  $y$ , a two-part coding scheme produces the following approximate formula for the MDL in terms of the optimal (MLE) estimator  $\hat{\theta}$ :

$$\text{MDL}(k) = -\log[p(y|\hat{\theta})\pi(\hat{\theta})] - (k \log n)/2 + O(k) \quad (48)$$

The last term is negligible for large  $n$  as is the  $\pi(\cdot)$  term.

Stochastic complexity for the same model class is obtained by eliminating a redundancy above that forces a non-essential parameter coding. Instead the following minimalist version of (48) is obtained as in chapter 2 of Rissanen (1989):

$$I(y|M_k) = -\log P(y) \quad (49)$$

$$P(y) = \int p(y|\theta) d\pi(\theta) \quad (50)$$

An approximate relationship between the MDL and stochastic complexity  $I(\cdot)$  is given in Rissanen (1989) as follows:

$$\begin{aligned} I(y|M_k) &\approx -\log P(y|\hat{\theta}) + \frac{1}{2} \log |\hat{\Sigma}| \\ &\approx -\log P(y|\hat{\theta}) + \frac{k}{2} \log n \end{aligned} \quad (51)$$

where  $\hat{\Sigma}$  denotes the matrix of the second derivative of the function  $L(y, \theta)$  with respect to  $\theta$ , and  $L(\cdot)$  is logarithm of the two-part code (joint likelihood). Interestingly, Rissanen shows that maximizing the posterior distribution does not provide a "meaningful criterion" especially since no penalty for the number of parameters will be given.

But information criteria have played a constructive role in Bayesian analysis, see Zellner (1991) for a recent survey. This role goes beyond the area of model selection on the basis of odds ratios. The ME principle can be used to examine "learning rules". Indeed Zellner and others have shown that Bayes' rule is an "information efficient" learning rule according to the Kullback-Leibler criterion. ME posterior distributions have also been derived by Zellner and his associates. The Kullback-Leibler criterion is sometimes referred to as the Jeffreys' criterion since he derived it in his search for invariant prior densities, see Jeffreys (1967).

## 5.2 Information Matrices and Metric Spaces

Let  $Y$  have the distribution function  $F(y, \theta)$  with a density denoted by  $f(y, \theta)$ . Let  $G$  be an hypothesized/competing distribution with density denoted by  $g(y, \theta)$ . And let  $h(\cdot)$  be an information function so that the entropy of  $F$  is as follows:

$$\begin{aligned}
 H_{\varphi}(f) &= \int h(\cdot) f dy \\
 &= \int \varphi(\cdot) dy, \text{ say, where } \varphi \text{ is known as the } \varphi\text{-entropy function.}
 \end{aligned}$$

It is then readily verified that:

$$-(d^2 H_{\varphi}(f)/df^2) = S_{\varphi}^2(\theta) = \sum_i \sum_j \lambda_{ij, \varphi} d\theta_i d\theta_j = d\theta \Lambda_{\varphi} d\theta \tag{52}$$

$$\lambda_{ij, \varphi} = -\int \varphi''(f) (\partial f / \partial \theta_i) (\partial f / \partial \theta_j) dy \tag{53}$$

$\Lambda_{\varphi}$  is called the  $\varphi$ -order information matrix, and  $S_{\varphi}^2$  is the  $\varphi$ -entropy metric (distance) of the Riemannian geometry. For Shannon's entropy,  $h(\cdot) = -\log f$ , and

$$\Lambda_{\varphi} = E [(\partial h / \partial \theta) (\partial h / \partial \theta)'] = \text{Fisher's information matrix.}$$

Note that  $S_{\varphi}^2(\cdot)$  is positive semi-definite for concave  $\varphi(\cdot)$ .

Burbea and Rao (1962) contains further elaboration. There is a fertile area of research on the differential geometry of statistics in general, and that of the Kullback-Leibler measure, in particular. Mahalanobis and C.R. Rao are amongst the earliest writers. See Critchley, Marriott, and Samon (1991) for a recent discussion.

**6. Axiom Systems for inequality and concentration**

In the remainder of this paper a sample of theoretical and empirical questions is given in which information theory concepts and tools have played an increasingly prominent and constructive role. The discussion focuses primarily on economic applications, but several other fields are at least as active in this area as economics.

**6.1 Welfare and Well-Being:**

One of the most perplexing questions of welfare analysis based on welfare and utility functions is what arguments should be included in these functions. For individuals, some candidates are income, wealth, health, education, and "needs". For social welfare functions, such arguments as individual utilities, entitlements, and liberty have been considered, especially since the UN and other agencies have procured data on other attributes of "well-being" than just "income". Accepting equality as a major component of equity, important advances have taken place in the last decade in developing the axiom systems which help in the classic problem of choice among inequality indices. In what follows it can be seen that these developments originate in

the literature on "ideal" information criteria and "ideal" functionals. Contributions of J. Aczel and other information theorists have provided the inspiration in this effort.

### 6.1.1 Welfarist/Utilitarian constructions

Let the Social Welfare Function be defined as follows:

$$SWF = W(u_1, u_2, u_3, \dots, u_N) \quad (1)$$

$u_i = u_i(x, y, z, \dots)$  stands for individual  $i$  "utility". What are the appropriate arguments in  $u_i$ ? in  $W$ ? We can include more than just commodities and services in  $u_i$  and stay "utilitarian", or go beyond  $u_i$ s in  $W$  and stay with "welfarism". The multi-dimensional approach as advocated by Sen, Atkinson & Bourguignon, Maasoumi and others, is a response to these required extensions. In the past income alone has been a dominant indicator of individual welfare, either as an argument of  $u_i$ , as  $u_i$ , or even as  $W$  itself! Whatever the attribute, the computation and analysis of the "inequality" in its distribution has been an essential element of welfare theory.

### 6.1.2 Measuring Income Inequality

Today there exist very many ad hoc measures of "equality" and several concepts of equality! Some measures are justified by convenience, some by statistical considerations, some have informal economic/welfarist appeal. Examples are Gini, Variance, Coefficient of Variation, Theil's Information Measures and numerous others. The axiomatic approach developed by Snorrocks, Bourguignon, Cowell and others, was a response to the need to rein in this arbitrary and confusing situation.

### 6.1.3 The Axiomatic Approach

Let the vector of incomes for  $N$  individuals be denoted by  $y = (y_1, y_2, \dots, y_N)$ , and any inequality measure by  $I(y)$ . The following "properties" or axioms are commonly required of any inequality index, as they would be of any information measure in a different guise and with different justifications:

A1 **Anonymity (symmetry)**:  $I(y)$  is invariant to permutations of  $y$ .

A2 **Normalization**:

$$\begin{aligned} &\text{If } y_i = \bar{y}, \text{ for all } i, \text{ then} \\ &I(\bar{y}) = 0. \quad \text{This is both innocuous and sensible.} \end{aligned}$$

**A3 Principle of Transfers (Pigou-Dalton):**

$$I(y') < I(y)$$

if  $y'$  is a "progressive" redistribution of  $y$ . i.e., if  $y_i > y_j$  for some  $i$  and  $j$ , redistribute  $y$  to  $y'$  such that  $y'_i = y_i - d > y_j + d = y'_j$ , and  $y_k = y'_k$  for all other  $k,s$ .

**A4 Continuity:**  $I(y)$  is continuous on the set of all income distributions  $F_N$ , for every  $N$ .

While A1-A4 are adequate for some "stochastic dominance" type comparisons (Lorenze), and rankings of distributions, there exist too many measures that satisfy them. These different inequality measures can also give conflicting rankings. So other less basic (but still sensible) properties are imposed:

**A5 Homogeneity:**  $I(y) = I(y')$ , if  $y = c y'$ , any scalar  $c$ . This will limit the class of admissible measures. Only "relative" inequality can be measured with indices of this type and "efficiency" is not addressed).

**A6 Replication Invariance:**

$I(y) = I(x)$  if  $x$  is a replication of  $y$ . E.g.,  $x = [y,y]$ . this is useful for dealing with populations of different sizes.

Still, axioms 1-6 are satisfied by many families of measures! Further restrictions and arguments are required, among which ethical and statistical considerations are prominent. An advantage of the formal axiomatic approach is that ethical values are made clear by connection to SWFs. But one also requires sensible properties relative to "usage", and/or logical practical (mathematical) considerations. But the SWF connection will not be enough. For example, let

$$W_1(y) = \sum_i u_i(y_i), \tag{55}$$

and use the following inequality index:

$$I(y) = 1 - y_e / \mu_y \tag{56}$$

where  $W(y) = W(y_e, y_e, \dots, y_e)$ , as a measure of relative welfare loss. This will produce, along with Axiom 5, the Atkinson family of measures.

This ethical connection, however, is not sufficient since we could as well generate a welfare function for any other choice of  $I(y)$ ! Take

$$W_2(y) = \mu_y \exp[-I(y)] \quad (57)$$

$$\text{and } I(y) = \ln(\mu_y) - \ln(y_e) \quad (58)$$

Additional "sensible" restrictions are desirable. e.g., if one wished to know "how much of the overall inequality in the world is due to the inequality among the Asian countries?", or "the centrally planned economies?", or "due to certain types of earnings such as social benefits?", or "due to sex, race or occupation differences?", then one would discover that most inequality indices are eliminated because they provide either vague or conflicting answers, see Shorrocks (1982). Hence additional restrictions resembling the "branching" axioms of information theory are needed. For example:

**A7 Aggregation Consistency (Additive decomposability)**

Let a population be made up of groups  $g=1,2,\dots,G$ , with group sizes  $n_g$ , exclusive, with mean incomes  $\bar{y}_g$ . Suppose we rearrange incomes within each subgroup (preserving  $\bar{y}_g$ ) such that the inequality in each group,  $I(y_g^B)$ , say, is increased. Then we would like the overall inequality for that population to increase. Surprisingly, such a sensible "subgroup consistency axiom" is violated by many indices including the most widely used of inequality indices, the Gini! This is unfortunate and raises serious questions regarding the folklore surrounding much empirical evidence. Thus we require:

$$I(y) = I(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_G) + \sum_g w_g I_g(y_g^B) \quad (59)$$

When combined with A1-A6 this requirement reduces the class of admissible measures to scalar multiples of the so-called **Generalized Entropy Indices**, see Shorrocks (1980,1984):

$$I_\gamma(y) = \sum_i [(y_i/\mu_y)^{\gamma+1} - 1] / N\gamma(\gamma+1), \quad \gamma \neq 0, -1 \quad (60)$$

$$I_0(y) = \sum_i (y_i/\mu_y) \ln(Ny_i/\mu_y), \quad (61)$$

is Theil's first measure, and

$$I_{-1}(y) = \frac{1}{N} \sum_i \ln(\mu_y/Ny_i) \quad (62)$$

is Theil's second measure when the "population share" of each unit is  $1/N$ . (60)-(62)

are clearly expected divergences between two distributions, the size distribution of income and the rectangular (uniform) distribution which has the highest entropy and represents "equality". When the population share of a typical unit is  $p_1$ , say, the above measures reflect the information divergence between the population and the income shares distributions.

Theil's second inequality index is known to allow for the least ambiguous additive decompositions into the within and between group components as follows:

$$I_{-1}(\bar{y}) = \sum \frac{n_g}{N} I(\bar{y}^g) + I(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_G), \quad (63)$$

Unlike other measures, particularly the Gini index, it is possible for a policy analyst to use Theil's measure and to isolate the contribution to the overall inequality of the inequality changes for a target group.

There is a common misunderstanding of the above measures as "entropy" measures. It would be a mistake to think that these measures involve either the concept or the computation of the entropy of the income distribution. In the above formulae the entropy of the population distribution is subtracted from another entropy related to incomes. But this latter entropy is different from the entropy of income distribution. It is in fact a mere entropy formula computed on the basis of income shares interpreted as probabilities. These are not the probabilities for the realizations of the income variable. Income shares define a "size distribution" whose entropy is computed for inequality measurement. It would be interesting to study the difference between the entropies of income and the population as a measure of inequality. Preliminary results suggest considerable distinction with the traditional measures, see Maasoumi (1992).

#### 7. Multidimensional Welfare and Inequality

Going beyond "income" and "utilitarianism" we may wish to consider, simultaneously, many welfare attributes such as in-kind "payments", social benefits, entitlements, freedoms (political freedoms, civil liberties), physical quality of life indices (PQLI, BN) and many of the other measured indicators of well-being. This

raises many new questions, such as measurement accuracy (relative to "hard" data), index number problems (see discussion below), the problem of "sensible axioms" or meaningful properties that aggregation functions and/or inequality measures should satisfy, and double counting problems.

Maasoumi (1979, 1980, 1990) developed an approach based on information theory. There are two steps in this approach. The first is to find an aggregate or composite measure of well-being. The second step is to apply a suitable measure of univariate inequality, say, to the aggregated measure. The choice of a measure in the second step will be guided by the axiomatic developments outlined above.

This approach has found applications in several areas, including the Michigan Panel study of income dynamics (PSID) data on all income sources, housing equity, and education, with grouping of the sample by age and sex; see Maasoumi and Nickelsburg (1988). There are other empirical applications in the areas of mobility, international "well-being" and inequality in GNP, PQLI, and BN, and in a "cluster analysis" study which attempts to identify "comparable" attributes in order to avoid double counting problems and over weighting [Hirschberg, Maasoumi, and Slottje (1992)]. A survey of empirical applications in inequality and welfare is provided in Maasoumi (1993).

#### 7.1 An Aggregation problem:

Take  $Y = [y_{ij}]$ , individual (unit)  $i=1, \dots, N$ ; attribute  $j=1, \dots, M$ . Let  $S_i = S_i(y_{i1}, y_{i2}, \dots, y_{iM})$ , be the aggregate function for the  $i$ -th unit. Find  $S_i$  such that  $S = (S_1, S_2, \dots, S_N)$ , is "closest" to all the  $M$ - attributes. The criterion of closeness, "comparability" of Sen's, is the Generalized or " $\gamma$ -Entropy":

$$GE(S) = \frac{1}{\gamma(\gamma+1)} \sum_j \alpha_j \left[ \sum_i S_i \left\{ (S_i/y_{ij})^\gamma - 1 \right\} \right] \quad (64)$$

This produces some surprising outcomes for "Ideal indices",  $S_i^*$ :

$$S_i^* \text{ - proportional to } \left[ \sum_j \delta_j y_{ij}^{-\gamma} \right]^{-1/\gamma} \quad (65)$$

where  $\delta_j$  are normalized  $\alpha_j$ . This Constant Elasticity of Substitution (CES) family includes such forms as the generalized geometric mean (Cobb-Douglas), linear, Leontief and other popular forms. Principal components have also been proposed as

composite indices, see Ram (1982). It can be seen that (65) includes principal components as a very special case obtained when  $\gamma = -1$  and  $\delta_j$ s are the elements of the first characteristic vector of the second moment matrix of  $y_{ij}$ . Viewed as a hyperbolic mean, (65) is also an example of solutions obtained for "ideal" averages in IT literature. Much of this work is done by Aczel and Roberts. Several empirical applications are described in section 8.

Maasoumi (1986a) proposed the axiomatic approach outlined above, and use of the GE measure  $I_\gamma(y)$  on the aggregate shares,  $S_i$ . Properties of this index as a multidimensional measure of inequality have been further studied in Maasoumi (1986b, 1989b) and Dardanoni (1992). Tsui (1992) has followed a one step, direct axiomatic approach to obtaining suitable aggregate functions and inequality measures on several attributes. Interestingly, essentially the same measures as in the two-step approach emerge!

In two recent papers, Maasoumi (1989a) and Maasoumi and Jeong (1985), information theoretic measures were utilized in the comparative analysis of well-being in the world. Traditional analysis in this area has hitherto been based almost exclusively on GNP or GDP. Typically, the size distribution of one of these attributes would be characterized by a Gini measure of inequality. This approach suffers from two principal shortcomings. The first is the use of Gini index, and the second is a more widely understood and accepted problem with the exclusive identification of well-being by pure income measures. The latter is particularly troubling in international comparisons where a good many significant other factors cannot be controlled amongst diverse attitudes, cultures, and socio-economic arrangements for individual and social attainment of welfare. Other relevant attributes have now been measured and made available by international agencies. Among the problems that must be faced, measurement accuracy, choice of "independent" attributes, how to combine the chosen attributes in composite indices, and the choice of an appropriate index of (e.g) inequality, may be mentioned. Information theory concepts were used in the above papers to deal with these problems except for measurement error.

The second difficulty, the almost exclusive reliance on the Gini measure of distributional characteristic, is also resolved if one uses the family of Generalized Entropy (GE) measures instead. Here economic theory and information theory together help to clarify a revealing correspondence between degrees of relative aversion to inequality (tail area transformations which are so important for policy analysis) and the choice of summary indices of inequality, poverty, etc (see section 7 above). Gini, being more sensitive to changes in the central area of distributions, is seen to be less appropriate than some entropy based measures for the study of dynamic distributional changes that may be expected as a result of typical development or tax-based policies.

In Maasoumi (1989a) per capita GNP was considered as well as two composite indices of Basic Needs (BN) indicators and Physical Quality of Life Indicators (PQLI). Each of the latter is made up of other components reflecting well-being. Two inequality indices, the entropy-based measures due to Theil, see Theil (1967), were computed for each of these three attributes for a distribution of countries in the middle to late 1970s. In addition, the first entropy based composite indices of GNP-BN and GNP-PQLI were proposed providing the first multi-dimensional measures of well-being based on information theory. This approach also provided an interpretation for some of the previously proposed composite measures, such as the Principal Components (PC) used by Ram (1982).

A similar application of these techniques to US panel data on household income, net worth in housing, and education was reported in Maasoumi and Nickelsburg (1989). The Michigan Panel Study of income dynamics provides a rich source of data on several thousand households, their demographic characteristics, their economic status in terms of income, education, equity, transfer payments, etc, starting in 1968. The above mentioned study focused on the two information measures of inequality proposed by Theil, for each of the three attributes, and for similar aggregates of them as in the international studies cited earlier.

In these applications, important "decomposition properties" of the information measures, described in a previous section in axiom A7, prove crucial in successfully

controlling for such group characteristics as economic system, geographical location, level of industrialization, and GNP level in the case of countries, and for levels of education, age, race, gender and income level, in the case of individuals and households in the second study.

A third set of applications of information criteria is the subject of Maasoumi and Zandvakili (1986,1989 and 1990). Whereas aggregation over different attributes was the focus of the previous studies above, aggregation over time of a single attribute, household income, is studied in this third area. Earlier, Shorrocks (1978) had considered simple sum of incomes over time and proposed a measure of "mobility". In contrast, the more general information aggregates of section 7, allow for more general substitution and weighting schemes over time. This can provide for more plausible means of controlling for transitory changes in attribute distributions, and provide mobility profiles as the aggregation interval is enlarged. Once again the decomposition/partition properties of the information measures allow us to control for such characteristics as race, education, age, sex and income level. In the absence of control for such "co-factors", the previous literature in this area has been unconvincing in its attribution of distributional changes and shifts to policy decisions and other socio-economic conditions.

Similar applications to the definition and measurement of industrial and/or trade "concentration" is possible and is of considerable value in economics and political science. Entropy, in its various forms, is a very natural and revealing measure of dispersion, but only Shannon entropy seems to have been noted and used in the industrial organization area; see Theil (1967) and Zandvakili (1992).

As the main theme of our survey suggests, anytime there is an interest in evaluating "similarity", "collapsibility", "proximity" and "closeness" among mathematical or statistical measurements, information criteria described earlier in this survey present themselves as, at least, very strong candidates. Cluster analysis is an important area in statistics generally, and social/economic science applications, in particular. Traditional measures of closeness used for clustering variables are rather

predictable variants of measures of "distance", such as the euclidean measure. See Hirschberg, Maasoumi and Slottje (1991) for a brief presentation of some of these measures. In the latter paper 24 attributes of "well-being" are considered for 120 countries. These attributes include GNP-related indicators as well as such variables as literacy rates, mortality rates, women's participation in the labor force, health status and infrastructure variables, political freedom and civil liberty indicators. As was mentioned above, "double counting" of similar attributes is a matter of concern when so many attributes are considered in any application, including ordinary regression analysis in the social sciences. Cluster analysis finds a natural application here, and in our study very interesting and plausible clusters of attributes were identified by traditional measures of closeness. Our measures of closeness did not include the information criteria that would sometime encompass the traditional criteria of "distance". This is remedied in Hirschberg, Maasoumi and Slottje (1992) where information criteria are added in evaluating the clustering characteristics of a large set of US welfare attributes in the last few decades. Some of the information measures used by these authors are the measures defined in section 2.3 above. It would be tempting to argue that the information criteria provide for a richer and more "informed" comparison of variables than squared distances, say. In the absence of a specific context this would be a reasonable conclusion from a consideration of the axiom systems that were exemplified in the first sections of this survey. But "information" and its value, or even quantity, needs to be judged in the context of the use to which it is being put. For example, all the information provided by a set of variables beyond their second moments is of questionable consequence in a linear regression model and to least squares based techniques of inference. So it is that the appropriate clustering criteria is at least partly decided by what use is to be made of the clustered groups of variables and data. This is a difficult issue, and we are aware of only some indirectly related results on this question in Maasoumi (1986a). More work in this area seems worthwhile.

In section 3, we alluded to the definition of financial entropy by M. Stutzer (1992). This paper contains an empirical application to daily data on US stocks. There is a growing number of empirical applications in finance and in conducting tests of dependence in non-linear dynamic models. Some limited progress has also been made in employing information theory concepts in order to quantify market "news" and "surprises".

#### References

- Aczel, J. and Z. Daroczy (1975) *On Measures of Information and Their Characterizations*, Academic Press, New York.
- Ahmad, I.A., and P.-I. Lin, (1976), "A nonparametric estimate of the entropy for absolutely continuous distributions," *IEEE Transactions on Information Theory*, IT-22, 372-375.
- Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," *Proc. of the second international symposium on information theory*, B.N. Petrov and F. Csaki, Akademiai Kiado, Budapest, 267-281.
- \_\_\_\_\_ (1977), "On entropy maximization principle," *Applications of Statistics*, Krishnaiah ed., North-Holland, 27-41.
- Amemiya, T. (1980), "Selection of regressors," *International Economic Review*, 21, 331-354.
- Andrews, D., and Y.-J. Whang (1990), "Additive interactive regression models: Circumvention of the curse of dimensionality," *Econometric Theory*, 6, 466-479.
- Atkinson, A.B., (1970), "On the measurement of inequality," *Journal of Economic Theory* 2, 244-263
- Atkinson, C. and A.F.S. Mitchell (1981), "Rao's distance measure," *Sankhya*, 43, series A, 3, 345-365.
- Begun, J., W. Hall, W. Huang, and J. Wellner, (1983), "Information and asymptotic efficiency in parametric-nonparametric models," *Annals of Statistics*, 11, 432-452.
- Bera, A.K., and P.T. Ng (1991), "Robust tests of heteroskedasticity and autocorrelation using score function," Working Paper, Univ. of Illinois.
- Bhattacharyya, A. (1943), "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, 35, 99-109.
- \_\_\_\_\_ (1946), "On a measure of divergence between two multinomial populations," *Sankhya*, 7, 401-406.
- Bickel, P.J. (1982), "The Walk Memorial lectures: On adaptive estimation," *Annals of Statistics*, 10, 647-671.

- Burbea, J. and C.R. Rao (1982), "Entropy differential metric, distance and divergence measures in probability spaces: A unified approach," *Journal of Multivariate Analysis*, 575-598.
- Chamberlain, G. (1987), "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.
- \_\_\_\_\_(1992), "Efficiency bounds for semiparametric regression," *Econometrica*, 60, 3, 567-596.
- Cover, T.M., and J.A. Thomas (1991), *Elements of information theory*, Wiley.
- Critchley, F., P. Marriott, and M. Samon (1991), "Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence," Working paper ECO No. 91/53, European University Institute, Florence.
- Csiszar, I., (1991), "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Annals of Statistics*, 19, 2032-2066.
- Dardanoni, Valentino, (1990), "A Note on Multidimensional Inequality Comparisons," University of California at San Diego, Working Paper.
- Davis, H.T., (1941) *The theory of econometrics*, The Principia Press, Bloomington, IN.
- Fisher, R., (1922) *The mathematical foundations of theoretical statistics*, *Philosophical Trans. of the Royal Statist. Soc. of London*, Series A, 222.
- Foster, J.E. (1983), "An Axiomatic Characterization of the Theil Measure of Income Inequality," *Journal of Economic Theory*, 31(1), 105-121.
- George, E.I. and R. McCulloch (1989), "On obtaining invariant prior distributions," Graduate school of business, Univ. of Chicago, November.
- Georgescu-Roegen, N. (1966), *Analytical Economics: Issues and Problems*, Harvard University Press, Boston.
- \_\_\_\_\_(1971), *The Entropy Law and the Economic Process*, Harvard University Press : Cambridge, MA.
- Good, I.J. (1963), "Maximum entropy for hypothesis formulation, especially on multidimensional contingency tables," *Annals of Mathematical Statistics*, 34, 911-934.
- Hartley, R.V.L. (1928), "Transmission of information," *Bell System Tech. Journal* 7, 379-423.
- Horvath, J. and F. Charvat (1967), "Quantification method of classification processes: concept of structural  $\alpha$ -entropy," *Kybernetika Cisle I. Rocnik*, 3, 30-34.
- Hirschberg, J., E. Maasoumi, and D.J. Slottje (1991), "Cluster Analysis and the Quality of Life Across Countries," *Journal of Econometrics*, Vol. 50, No.1/2, pp. 131-150.
- \_\_\_\_\_(1992), "A dynamic analysis of well-being in the US based on clusters of attributes," Dept. of Economics, SMU, Dallas, Texas.

- Huber, P.J. (1984), "Robust estimation of a location parameter," *Annals of Math. Stats.*, 35, 73-101.
- Jaynes, E., (1979), "Concentration of distributions," in R. Rosenkrantz (ed.), *E. Jaynes : Papers on Probability, Statistics and Statistical Physics*, Reidel, Dordrecht.
- Jeffreys, H. (1967), *Theory of probability*, (3rd. rev. ed.), Oxford University Press, London.
- Joe, H. (1989), "Relative entropy measures of multivariate dependence," *JASA*, 84, 157-164.
- Joiner, B. and D. Hall (1983), "The ubiquitous role of  $I'/I$  in efficient estimation of location," *The American Statistician*, 37, 128-133.
- Jones, L.K. and Byrne, C.L. (1990), "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Trans. On Info. Theory*, 36, 1, 23-30.
- Judge, G., W. Griffiths, C. Hill, H. Lutkepohl, and T-C. Lee (1985) *The theory and practice of econometrics*, Second ed., Wiley.
- Kallianpour, G. (1960), "On the amount of information in a sigma field," in I. Olkin et al. eds., *Contributions to Probability and Statistics in Honor of H. Hotelling*, Stanford U. Press.
- Khinchin, A.I. (1957), *Mathematical foundations of information theory*, Dover Pub.
- Kirman, S.N.U.A. (1979), "On the relation between Matusita's and Kolmogorov's measures of distance," *Annals of Institute of Statist. Math.*, 31, Part A, 289-291.
- Klein, R.W., and S.J. Brown (1989), "Model selection under "minimal" prior information," Bell Labs., Murray Hill.
- Koenker, R. (1982), "Robust methods in econometrics," *Econometric Reviews*, 1, 214-255.
- Kolmogorov, A.N. (1963), "On the approximation of distribution of sums of independent summands by infinitely divisible distributions," *Sankya, A*, 25, 159-179.
- Kullback, S. (1959), *Information theory and statistics*, New York: Wiley.
- Kullback, S. and R.A. Leibler (1951), "On Information and sufficiency," *Ann. of Math. Stats.*, 22, 79-86.
- Leamer, E.E. (1979), "Information criteria for the choice of regression models, a comment," *Econometrica*, 47, 507-510.
- Maasoumi, E. (1979), "A measure of multivariate inequality," Mimeo., University of Southern California.
- \_\_\_\_\_ (1985), "Unknown Regression Functions and Information Criteria," Indiana University, Department of Economics.
- \_\_\_\_\_ (1986a), "The measurement and decomposition of multidimensional inequality," *Econometrica*, 54, 991-997.

- \_\_\_\_\_ (1986b), "Unknown regression functions and information efficient functional forms: An interpretation," *Advances in Econometrics*, Vol.5, 301-309.
- \_\_\_\_\_ (1988a), "On econometric methodology," *Economic Record*, December.
- \_\_\_\_\_ (1988b), "Information Theory," in *The New Palgrave: A Dictionary of Economics*, Vol.2, New York: Stockton Press, 846-51; Reprinted in *New Palgrave: Econometrics*, Norton, 1990.
- Maasoumi, E. (1989a), "Composite Indices of Income and Other Developmental Indicators: A General Approach," *Research on Economic Inequality*, Vol.1, 269-286.
- \_\_\_\_\_ (1989b), "Continuously Distributed Attributes and Measure of Multivariate Inequality," *Journal of Econometrics*, 131-144.
- \_\_\_\_\_ (1992), "A Clarification of Entropy as a Measure of Inequality," Department of Economics, SMU.
- \_\_\_\_\_ (1993), "Empirical studies of well-being and inequality," *Handbook of Applied Microeconometrics*, M.H. Pesaran and P. Schmidt (eds.), in preparation.
- \_\_\_\_\_ and Jin-Ho Jeong (1985), "The Trend and the Measurement of World Inequality Over Extended Periods of Accounting," *Economics Letters*, 19, pp. 295-301.
- \_\_\_\_\_ and G. Nickelsburg (1988), "Multivariate Measures of Well-Being and an Analysis of Inequality in the Michigan Data," *Journal of Business and Economic Statistics*, Vol. 6, 3, 327-334.
- \_\_\_\_\_ and H. Theil (1979), "The Effect of the Shape of the Income Distribution on Two Inequality Measures," *Economics Letters*, 4, pp.289-291.
- \_\_\_\_\_ and S. Zandvakili (1986), "A Class of Generalized Measures of Mobility with Applications," *Economics Letters*, 97-102.
- \_\_\_\_\_ (1989), "Mobility Profiles and Time Aggregates of Individual Incomes," *Research on Economic Inequality*, Vol 1, 195-218.
- \_\_\_\_\_ (1990), "Generalized Entropy Measures of Mobility in Different Sexes and Income Levels," *Journal of Econometrics*, 121-133.
- Maasoumi, E. and Zhang, G. (1992a), "Generalized entropy measures of volatility in US stock returns," Dept. of Economics, SMU (Mimeographed).
- \_\_\_\_\_ (1992b), "Numerical algorithms for generalized entropy of distribution functions," Dept. of Economics, SMU (mimeo.)
- Mahanalobis, P.C. (1930), "On the test and measures of group divergences," *Journal and Proceedings of the Asiatic Society of Bengal*, New series, 26, No.4, 541-588.
- Mallows, C.L. (1973), "Some comments on  $C_p$ ," *Technometrics*, 15, 661-676.

- Manski, C. (1984), "Adaptive estimation of non-linear regression models," *Econometric Reviews*, 3, 145-194.
- Matusita, K. (1951), "On the theory of decision functions," *Ann. Inst. Statist. Math.*, 3, 17-35.
- \_\_\_\_\_ (1967), "On the notion of affinity of several distributions and some of its applications," *Ann. Inst. Statist. Math.*, 19, 181-192.
- Newey, W. (1988), "Adaptive estimation of regression model via moment restrictions," *Journal of Econometrics*, 38, 301-339.
- Pagan, A.R. and A.D. Hall (1983), "Diagnostic tests as residual analysis," *Econometric Reviews*, Vol.2, 2, 159-254.
- Parzen, E. (1982), "Maximum entropy interpretation of autoregressive spectral densities," *Stats. and Prob. Letters*, 1, 2-8.
- Parzen, E. (1990), "Time series, statistics, and information," IMA preprint series # 663
- Pinsker, M.S., (1964) **Information and information stability of random variables and processes**, San Francisco: Holden-day.
- Prescott, P. (1976), "On a test of Normality based on sample entropy," *Journal of the Royal Statistical Society*, B, 38, 254-256.
- Rao, C.R. (1945), "Information and the accuracy obtainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, 37, 81-91.
- Ram, R., (1982), "Composite indices of physical quality of life, basic needs fulfillment, and income: A 'principal component' representation," *Journal of Development Economics*, 11, 227-247
- Renyi, A. (1961), "On measures of entropy and information," *Proc. 4th Berkeley Symposium. Statist. Probability*, 1, 547-561; U. of California Press.
- \_\_\_\_\_ (1967), "Statistics and Information Theory," *Studia Sci. Math. Hungarica* 2, 249-256.
- Rissanen, J. (1987a), "Stochastic complexity," *J. of Royal Statistical Society, Series B*, 49, No.3, (with discussion), 223-265.
- \_\_\_\_\_ (1987b), "Stochastic complexity and the MDL principle," *Econometric Reviews*, Vol. 6, 1, 85-102.
- \_\_\_\_\_ (1989), **Stochastic Complexity in Statistical Inquiry**, Vol. 15, World Scientific, Series in computer science.
- Robinson, P.M. (1991), "Consistent nonparametric entropy-based testing," *Review of Economic Studies*, 58, 437-453.
- Ryu, H.K., (1991), "Maximum entropy estimation of density and regression functions," *Journal of Econometrics*, forthcoming.
- Sawa, T. (1978), "Information criteria for discriminating among alternative regression models," *Econometrica*, 46, 1273-1291.

Shannon, C.E. (1948), "The mathematical theory of communication," *Bell System Tech Journal*, 27, 379-423, 623-656, and in Shannon and Weaver (1949), *The mathematical theory of communication*, Univ. of Illinois, Urbana, 3-91.

Shore, J. and R. Johnson (1980), "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Information Theory*, IT-26, 26-37.

Shorrocks, A.F. (1978), "Income Inequality and Income Mobility," *Journal of Economic Theory*, 19, 376-393.

\_\_\_\_\_. (1980), "The class of additively decomposable inequality measures," *Econometrica*, 48, 613-625.

\_\_\_\_\_. (1984), "Inequality decompositions by population subgroups," *Econometrica*, 52, 1369-1385.

Sin, C-Y. and H. White (1992), "Information criteria for selecting possibly misspecified parametric models," Department of Economics, UCSD working paper (November).

Soofi, E. S. (1990), "Effects of collinearity on information about regression coefficients," *Journal of econometrics*, 255-274.

\_\_\_\_\_, and D.V. Gokhale (1991), "An information criterion for normal regression estimation," *Statistics and Probability Letters*, 111-117.

Stutzer, M. (1992), "Arbistatics", Dept. of Finance, Univ. of Minnesota (December).

Theil, H. (1967) *Economics and Information Theory*, Rand McNally, Chicago, Illinois.

Theil, H. and D. C. Fiebig (1984) *Exploiting continuity: Maximum Entropy Estimation of Continuous Distributions*, Ballinger, Cambridge, MA.

Theil, H. and C. Laitinen (1980), "Singular moment matrices in applied econometrics," in P.R. Krishnaiah (ed.) *Multivariate Analysis - V*, North-Holland publishing Co, 629-649.

Tintner, G. (1960), "Applications of the theory of information to the problem of weighted regression," *Onore De Corrado Gini*, 1, 29; Rome Inst. De Statist. University;

Tsui, K-Y. (1992), "Multidimensional Generalizations of the Relative and Absolute Inequality Indices: The Atkinson-Kolm-Sen Approach," Chinese University of Hong Kong, Shatin, Hong Kong, Working Paper.

Vasicek, O. (1976), "A test for normality based on sample entropy," *Journal of Royal Statistical Society*, 38B, 54-59.

Weiner, N. (1949) *Cybernetics*, New York, Wiley.

White, H. (1990), "A consistent model selection," in C.W.J. Granger (ed.) *Modelling Economic Series*, 369-383, Oxford University Press: Oxford.

\_\_\_\_\_. (1992) *Estimation, inference and specification analysis*, Cambridge U. Press, forthcoming.

Whittle, P. (1953), "Estimating and information in stationary time series," *Ark. Math.*, 2, 423-434.

Zandvakili, S. (1992), "Multi-Attribute, Multi-Period Measurement of Industrial Concentration: An Information Theoretic Approach," University of Cincinnati, Economics.

Zelner, A. (1991), "Bayesian methods and entropy in economics and econometrics," In W.T. Grandy and L.H. Schick (eds), *Maximum Entropy and Bayesian Methods*, 17-31, Kluwer.

Zelner, A., and R. Highfield (1988), "Calculation of maximum entropy distributions and approximation of marginal distributions," *Journal of Econometrics*, 37, 195-209.