Information and Entropy Econometrics – Editor's View

Amos Golan*

Department of Economics, American University, Roper 200, 4400 Massachusetts Ave., NW, Washington, DC 20016, USA.

1. Introduction

Information and Entropy Econometrics (IEE) is research that directly or indirectly builds on the foundations of Information Theory (IT) and the principle of Maximum Entropy (ME). IEE includes research dealing with statistical inference of economic problems given incomplete knowledge or data, as well as research dealing with the analysis, diagnostics and statistical properties of information measures. By understanding the evolution of ME, we can shed some light on the roots of IEE.

The development of ME occurred via two lines of research:

i) The 18th century work (principle of insufficient reason) of Jakob Bernoulli (published eight years after his death, 1713)[1], Bayes (1763) and Laplace (1774): They all investigated the basic problem of calculating the state of a system based on a limited number of expectation values (moments) represented by the data. This work was later generalized by Jeffreys (1939) and Cox (1946). This line of research is known as Statistical Inference.

ii) The 19th century work of Maxwell (1859, 1876) and Boltzmann (1871), continued by Gibbs (1902) and Shannon (1948): This work is geared toward developing the mathematical tools for statistical modeling of problems in mechanics, physics and information.

The two independent lines of research are similar. The objective of the first line of research is to formulate a theory/methodology that allows understanding of the general characteristics (distribution) of a system from partial and incomplete information. In the

* Corresponding author. Tel.: 001-202-885-3783; Fax: 001-202-885-3790.
E-mail address: agolan@american.edu (A.Golan)

[1] Note that Jakob Bernoulli is also known as Jacque and James Bernoulli.

second line of research, this same objective is expressed as determining how to assign (initial) numerical values of probabilities when only some (theoretical) limited global quantities of the investigated system are known. Recognizing the common basic objectives of these two lines of research aided Jaynes (1957) in the development of his classical work, the Maximum Entropy (ME) formalism. The ME formalism is based on the philosophy of the first line of research (Bernoulli, Bayes, Laplace, Jeffreys, Cox) and the mathematics of the second line of research (Maxwell, Boltzmann, Gibbs, Shannon).

The interrelationship between Information Theory (IT), statistics and inference, and the ME principle started to become clear in the early work of Kullback, Leibler and Lindley. Building on the basic concepts and properties of IT, Kullback and Leibler developed some of the fundamental statistics, such as sufficiency and efficiency as well as a generalization of the Cramer-Rao inequality, and thus were able to unify heterogeneous statistical procedures via the concepts of IT (Kullback and Leibler 1951; Kullback 1954, 1959). Lindley (1956), on the other hand, provided the interpretation that a statistical sample could be viewed as a noisy channel (Shannon's terminology) that conveys a message about a parameter (or a set of parameters) with a certain prior distribution. In that way, he was able to apply Shannon's ideas to statistical theory by referring to the information in an experiment rather than in a message.[2]

The interrelationship between Information Theory (IT), statistics and inference, and the ME principle may seem at first as coincidental and of interest only in a small number of specialized applications. But, by now it is clear that when these methods are used in conjunction, they are useful for analyzing a wide variety of problems in most

[2] For a nice detailed discussion see Soofi (1994).

disciplines of science. Examples include (i) work on image reconstruction and spectral analysis in medicine, physics, chemistry, biology, topography, engineering, communication and information, operations research, political science and economics (e.g., brain scan, tomography, satellite images, search engines, political surveys, input-output reconstruction and general matrix balancing), (ii) research in statistical inference and estimation, and (iii) ongoing innovations in information processing and IT.

The basic research objective of how to formulate a theory/methodology that allows understanding of the general characteristics (distribution) of a system from partial and incomplete information has generated a wide variety of theoretical and empirical research. That objective may be couched in the terminology of statistical decision theory and inference in which we have to decide on the "best" way of reconstructing an image (or a "message" in Shannon's work), making use of partial information about that image. Similarly, that objective may be couched within the more traditional terminology, where the basic question is how to recover the most conservative estimates of some unknown function from limited data. The classical ME is designed to handle such questions and is commonly used as a method of estimating a probability distribution from an insufficient number of moments representing the only available information.

IEE is a natural continuation of IT and ME. All of the studies in IEE (developed mostly during the 1990s) build on both IT and/or ME to better understand the data while abstracting away from distributional assumptions or assumptions on the likelihood function. The outcome of these independent lines of study was a class of information-based estimation rules that differ but are related to each other. All of these types of

methods perform well and are quite applicable to large classes of problems in the natural sciences and social sciences in general, and in economics in particular.

The objectives of this volume are to gather a collection of articles from the wide spectrum of topics in IEE and to connect these papers and research together via its natural unified foundation: ME and IT. To achieve these objectives, the papers in this volume include summaries, reviews, state of the art methods, as well as discussion of possible future research directions.

2. Brief Summary of Recent History

2.1. Information and Entropy - Background

Let $\boldsymbol{A} = \{a_1, a_2, ..., a_M\}$ be a finite set and $\boldsymbol{p}$ be a proper probability mass function on $\boldsymbol{A}$. The amount of information needed to fully characterize all of the elements of this set consisting of *M* discrete elements is defined by $I(\boldsymbol{A}_M) = \log_2 M$ and is known as Hartley's formula. Shannon (1948) built on Hartley's formula, within the context of communication process, to develop his information criterion. His criterion, called entropy,[3] is

$$H(\boldsymbol{p}) \equiv -\sum_{i=1}^{M} p_i \log p_i \qquad (2.1a)$$

with *xlog(x)* tending to zero as *x* tends to zero. This information criterion measures the uncertainty or informational content that is implied by $\boldsymbol{p}$. The entropy-uncertainty measure H($\boldsymbol{p}$) reaches a maximum when $p_1 = p_2 = ... = p_M = 1/M$ (and is equal to Hartley's formula) and a minimum with a point mass function. It is emphasized here that

[3] In completing his work, Shannon noted that "information" is already an overused term. He approached his colleague John von Newman, who responded: "You should call it entropy for two reasons: first, the function is already in use in thermodynamics under the same name; second, and more importantly, most people don't know what entropy really is, and if you use the word *entropy* in an argument you will win every time".

H(***p***) is a function of the probability distribution. For example, if $\boldsymbol{h}$ is a random variable with possible distinct realizations $x_1, x_2, ..., x_M$ with probabilities $p_1, p_2, ..., p_M$, the entropy H(***p***) does not depend on the values $x_1, x_2, ..., x_M$ of $\boldsymbol{h}$. If, on the other hand, $\boldsymbol{h}$ is a continuous random variable, then the (differential) entropy of a continuous density is

$$H(\boldsymbol{X}) \equiv -\int p(x) \log p(x) dx \qquad (2.1b)$$

where this differential entropy does not have all of the properties of the discrete entropy (2.1a). For a further detailed and clear discussion of the entropy concept and of Information Theory see Cover and Thomas (1991) and Soofi (1994).

After Shannon introduced this measure, a fundamental question arose: whose information does this measure capture? Is it the information of the "sender", the "receiver" or the communication channel?[4] To try and answer this question, let us first suppose that *H* measures the state of ignorance of the receiver that is reduced by the receipt of the message. But this seemingly natural interpretation contradicts Shannon's idea. He used *H* to measure the overall capacity required in a channel to transmit a certain message at a given rate. Therefore, *H* is free of the receiver's level of ignorance. So what does it measure?

One answer to this question is that *H* is a measure of the amo unt of information in a message. To measure information, one must abstract away from any form or content of the message itself. For example, in the old-time telegraph office, where only the number of words were counted in calculating the price of a telegram, one's objective was to minimize the number of words in a message while conveying all necessary information.

---

[4] Within the context of IT, "channel" means any process capable of transmitting information.

Likewise, the information in a message can be expressed as the number of signs (or distinct symbols) necessary to express that message in the most concise and efficient way. Any system of signs can be used, but the most reasonable one is to express the amount of information by the number of signs necessary to express it by zeros and ones. In that way, messages and data can be compared by their informational content. Each digit takes on the values 0 or 1, and the information specifying which of these two possibilities occurred is called a unit of information. The answer to a question that can only be answered by "yes" and no" contains exactly one unit of information regardless of the meaning of that question. This unit of information is called a "bit" or binary digit.[5] Further, Renyi (1961, 1970) showed that, for a (sufficiently often) repeated experiment, one needs on average the amount $H(\boldsymbol{p})+\boldsymbol{e}$ of zero-one symbols (for any positive $\boldsymbol{e}$) in order to characterize an outcome of that experiment. Thus, it seems logical to claim that the outcome of an experiment contains the amount of information $H(\boldsymbol{p})$.

The information discussed here is not "subjective" information of a particular researcher. The information observed in a single observation, or a data set, is a certain quantity that is independent of whether the observer (e.g., an economist or a computer) recognizes it or not. Thus, $H(\boldsymbol{p})$ is a measure of the average amount of information provided by an outcome of a random drawing governed by $\boldsymbol{p}$. Similarly, $H(\boldsymbol{p})$ is a measure of uncertainty about a specific possible outcome before observing it, which is equivalent to the amount of randomness represented by $\boldsymbol{p}$.

---

[5] Shannon's realization that the binary digits could be used to represent words, sounds, images and ideas, is based on the work of George Boole, the 19th-century British mathematician, who invented the two-symbol logic in his work "The Laws of Thought."

According to both Shannon and Jaynes (1957), *H* measures the degree of ignorance of a communication engineer who designs the technical equipment of a communication channel because it takes into account the set of all possible messages to be transmitted over this channel during its life time. In more common econometric terminology, we can think of *H* in the following way. The researcher never knows the true underlying values characterizing an economic system. Therefore, one may incorporate her/his understanding and knowledge of the system in constructing the image where this knowledge appears in terms of some global quantities, such as moments. Out of all possible such images, where these moment conditions are retained, one should choose the image having the maximum level of entropy. The entropy of the analyzed economic system is a measure of the ignorance of the researcher who knows only some moments' values representing the underlying population. For a more detailed discussion of the statistical meaning of information see Renyi (1970) and Soofi and Retzer (this volume).

If, in addition, some prior information $\boldsymbol{q}$, defined on $\boldsymbol{A}$, exists, the cross-entropy (or Kullback-Leibler, K-L, 1951) measure is $I(\boldsymbol{p};\boldsymbol{q}) = \sum_{i=1}^{M} p_i \log(p_i / q_i)$ where a uniform $\boldsymbol{q}$ reduces $I(\boldsymbol{p};\boldsymbol{q})$ to $H(\boldsymbol{p})$. This measure reflects the gain in information with respect to $\boldsymbol{A}$ resulting from the additional knowledge in $\boldsymbol{p}$ relative to $\boldsymbol{q}$. Like with $H(\boldsymbol{p})$, $I(\boldsymbol{p};\boldsymbol{q})$ is an information-theoretic distance of $\boldsymbol{p}$ from $\boldsymbol{q}$. For example, if Ben believes the random drawing is governed by $\boldsymbol{q}$ (for example, $q_i = 1/M$ for all $i=1, 2, \ldots, M$) while Maureen knows the true probability $\boldsymbol{p}$ (which is different than uniform), then $I(\boldsymbol{p};\boldsymbol{q})$ measures how much less informed Ben is relative to Maureen about the possible outcome. Similarly, $I(\boldsymbol{p};\boldsymbol{q})$ measures the gain in information when Ben learns that

Maureen is correct – the true distribution is $\boldsymbol{p}$, rather than $\boldsymbol{q}$. Phrased differently, $I(\boldsymbol{p};\boldsymbol{q})$ may also represent loss of information such as Ben's loss when he uses $\boldsymbol{q}$. For further discussion of this measure see Cover and Thomas (1991), Maasoumi (1993) and Soofi and Retzer (this volume).

2.2. Maximum Entropy - Background

Facing the fundamental question of drawing inferences from limited and insufficient data, Jaynes proposed the ME principle, which he viewed as a generalization of Bernoulli and Laplace's Principle of Insufficient Reason. Using the tools of the calculus of variations the classical ME is briefly summarized.[6]

Given T structural constraints in the form of moments of the data (distribution), Jaynes proposed the ME method, which is to maximize $H(\boldsymbol{p})$ subject to the T structural constraints. Thus, if we have partial information in the form of some moment conditions, X_t $(t=1, 2, \ldots, T)$, where $T<M$, the ME principle prescribes choosing the $p(a_i)$ that maximizes $H(\boldsymbol{p})$ subject to the given constraints (moments) of the problem. These constraints can be viewed as certain "conservation laws" or "moment conditions" that represent the available information. His solution to this underdetermined problem is

$$\hat{p}(a_i) \propto \exp\left\{-\sum_t \hat{l}_t X_t(a_i)\right\} \tag{2.2}$$

where $\boldsymbol{l}$ are the T Lagrange multipliers, and $\hat{\boldsymbol{l}}$ are the values of the optimal solution (estimated values) of $\boldsymbol{l}$. Naturally, if no constraints (data) are imposed, $H(\boldsymbol{p})$ reaches its maximum value and the p's are distributed uniformly.

Specifically, if the available information is in the form of

$$\sum_i p_i = 1 \text{ and } \sum_i p_i g_t(X_i) = E[g_t],\ t=1, 2, \ldots, T, \tag{2.3}$$

where E is the expectation operator and $g_0(X_i) \equiv 1$ for all i, then the least "informed" (prejudiced) proper distribution that obeys these $T+1$ restrictions is:

$$\hat{p} = \exp\left\{-\hat{l}_0 - \hat{l}_1 g_1(X_i) - \hat{l}_2 g_2(X_i) \cdots - \hat{l}_T g_T(X_i)\right\} = \exp\left\{-\sum_{t=0}^{T} \hat{l}_t g_t(X_i)\right\}. \tag{2.4}$$

The entropy level is

$$H = \hat{l}_0 + \sum_{t=1}^{T} \hat{l}_t E[g(X_i)]. \tag{2.5}$$

The partition function (known also as normalization factor or the potential function), l_0, is defined as

$$l_0 = \log\left[\sum_i \exp\left(-\sum_{t=1}^{T} \hat{l}_t g_t(X_i)\right)\right] \tag{2.6}$$

and the relationship between the Lagrange multipliers and the data is given by

$$-\frac{\partial l_0}{\partial l_t} = E[g_t] \tag{2.7}$$

while the higher moments are captured by

$$\frac{\partial^2 l_0}{\partial l_t^2} = Var(g_t) \text{ and } \frac{\partial^2 l_0}{\partial l_t \partial l_s} = Cov(g_t g_s). \tag{2.8}$$

With that basic formulation, Jaynes was able to "resolve" the debate on probabilities vs. frequencies by *defining* the notion of probabilities via Shannon's entropy measure. His principle states that in any inference problem, the probabilities should be

[6] ME is a standard variational problem. See for example, Goldstine (1980) and Sagan (1993).

assigned by the ME principle, which maximizes the entropy subject to the requirement of proper probabilities and any other available information.[7]

Prior knowledge can be incorporated into the ME framework by minimizing the cross-entropy, rather than maximizing the entropy, subject to the observed moments. If $\tilde{\mathbf{p}}$ is the solution to such an optimization problem, then it can be shown that $I(\mathbf{p};\mathbf{q}) = I(\mathbf{p};\tilde{\mathbf{p}}) + I(\tilde{\mathbf{p}};\mathbf{p})$ *for any* $\mathbf{p}$ *satisfying the set of constraints (2.3),* which is the analogous to the Pythagorean Theorem in Euclidean geometry, where $I(\mathbf{p};\mathbf{q})$ can be regarded as the analogous for the squared Euclidean distance.

There exists an important interpretation of (2.4)-(2.6) within the context of Bayes theorem. The exact connection between ME, information and Bayes theorem is developed in Zellner (1988), discussed in the various papers of Jaynes, and is generalized in this volume (Zellner, 2001).

Finally, one cannot ignore two basic questions that keep coming up: Is the ME principle "too simple?" and does the ME principle "produce something from nothing?" The answer to the above is contained in the simple explanation that this principle uses only the relevant information, and eliminates all irrelevant details from the calculations by averaging over them.

[7] In the fields of economics and econometrics, it was probably Davis (1941) who conducted the first work within the spirit of ME. He conducted this work before the work of Shannon and Jaynes, and therefore he did not use the terminology of IT/ME. In his work, he estimated the income distribution by (implicitly) maximizing the Stirling's approximation of the multiplicity factor subject to some basic requirements/rules. A nice discussion of his work and of the earlier applications and empirical work in IEE in general and ME in particular in economics/econometrics appears in Zellner (1991) and Maasoumi (1993). For recent theoretical and applied work see the papers and citations provided in the current volume.

2.3. Information, Entropy and Maximum-Entropy Revisited

Building on Shannon's work, a number of generalized information measures were developed. Starting with the idea of describing the gain of information, Renyi (1961) developed the entropy of order α for incomplete random variables.[8] The relevant generalized entropy measure of a *proper* probability distribution (Renyi, 1970) is

$$H_a^R(\boldsymbol{p}) = \frac{1}{1-\boldsymbol{a}} \log \sum_k p_k^a \ . \tag{2.9}$$

The Shannon measure is a special case of this measure where $\boldsymbol{a} \to 1$. Similarly, the (Renyi) cross entropy of order α is

$$I_a^R(\boldsymbol{x} \mid \boldsymbol{y}) = I_a^R(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{1-\boldsymbol{a}} \log \sum_k \frac{p_k^a}{q_k^{a-1}}, \tag{2.10}$$

which is equal to the traditional cross-entropy measure as $\boldsymbol{a} \to 1$.

Building on Renyi's work, and independent of his work, a number of other generalizations were developed. These generalizations include the less known Bergman distance and the f-entropy measures. However, the commonly used generalized measures in IEE are those that were developed during the 1980's by Cressie and Read (1984) and Tsallis (1988). The cross-entropy version of the Tsallis measure is

$$I_a^T(\boldsymbol{x} \mid \boldsymbol{y}) = I_a^T(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{1-\boldsymbol{a}} \left(\sum_k \frac{p_k^a}{q_k^{a-1}} - 1 \right), \tag{2.11}$$

and the commonly used Cressi-Read measure is

[8] If $\boldsymbol{h}$ is an incomplete random variable with M distinct realizations, then $\sum_i p_i \le 1$ (rather than $\sum_i p_i = 1$) where $p_i > 0$; *i=1, …, M.*

$$I_a^{CR}(\boldsymbol{x} \mid \boldsymbol{y}) = I_a^{CR}(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{\boldsymbol{a}(1+\boldsymbol{a})} \sum_k p_k \left[\left(\frac{p_k}{q_k} \right)^{\boldsymbol{a}} - 1 \right]. \qquad (2.12)$$

Although it has not been recognized in the literature, we can show that all of these measures are connected. To do so, we compare the Tsallis and Renyi measures of order (α+1) with that of Cressi-Read of order α:

$$I_{a+1}^{R}(\boldsymbol{p}, \boldsymbol{q}) = -\frac{1}{\boldsymbol{a}} \log[1 - \boldsymbol{a} I_{a+1}^{T}(\boldsymbol{p}, \boldsymbol{q})] = -\frac{1}{\boldsymbol{a}} \log[1 + \boldsymbol{a}(\boldsymbol{a}+1) I_a^{CR}(\boldsymbol{p}, \boldsymbol{q})] \qquad (2.13)$$

where the traditional cross-entropy measure is a special case of the above for $\boldsymbol{a} \to 0$. (Since the other entropy measures are not of interest here, their connection with the above measures is not discussed here.) All of the above measures are commonly known as $\boldsymbol{a}$-*entropies*. For completeness, we note that the α-entropy is also known as "Chernoff entropy." Chernoff (1952) introduced this measure in his classical work on asymptotic efficiency of hypothesis tests. Chernoff entropy is found by starting with (2.13), and letting $\boldsymbol{a}=1-\boldsymbol{b}$ with $0<\boldsymbol{b}<1$. For the basic properties of these measures see Golan and Perloff (this volume).

All of the estimation methods within IEE are based on optimizing Eq. 2.13 for given values of $\boldsymbol{a}$*, subject to certain moment representation of the data, or certain "conservation laws" representing the underlying system.* This class of methods is led by the pioneering work in the Bayesian Method of Moments (BMOM), the Empirical Likelihood (EL), variations of the Generalized Method of Moments (GMM) and the Generalized ME (GME). All of these methods share the same basic objective of analyzing limited and noisy data using minimal assumptions.

Specifically, econometricians are often faced with finite and non-experimental data sets that in most cases are ill behaved. Further, as the underlying data generating process (or error margins) is uncertain or unknown, statisticians and econometricians try to avoid strong distributional assumptions or a pre-specified likelihood function. With the above in mind, and within the general objective of estimation and inference for a large class of models (linear and nonlinear, parametric and non-parametric), it seems that going back to the foundations of IT and ME was quite inevitable and led to a whole class of information-theoretic methods. All of these information-theoretic methods could be viewed as approaches to solving ill-posed or under-determined problems in the sense that without a pre-specified likelihood or distribution, there are always more unknowns than knowns regardless of the amount of data. That is, since the observation matrix is irregular or ill-conditioned or since the number of unknowns exceeds the number of data points, the problem is ill-posed. To solve these problems, one has to (i) incorporate some prior knowledge, or constraints, on the solution, or (ii) specify a certain criterion to choose among the infinitely many solutions, or (iii) use both approaches. But what criterion and what constraints should one use?

It seems natural to employ an informational criterion together with variations of the observed moments. For example, Zellner (1997, p. 86) says, "The BMOM approach is particularly useful when there is difficulty in formulating an appropriate likelihood function. Without a likelihood function, it is not possible to pursue traditional likelihood and Bayesian approaches to estimation and testing. Using a few simple assumptions, the BMOM approach permits calculation of post-data means, variances and other moments of parameters and future observations."

In the BMOM approach, one starts by maximizing the continuous entropy function (the continuous version of Eq. 2.13 with $\alpha \to 0$) subject to some *T* side conditions (pure conservation laws) and normalization. This approach yields the average log-height of the density function, which is the least informative density given these side-conditions.

Similarly, under the EL objective, one starts by searching for the "natural" weight of each observation by maximizing Eq. 2.13 (with $\alpha \to -1$) subject to the exact moment restrictions and normalization. Under the GME approach, one maximizes 2.13 (with $\alpha \to 0$ and with respect to both signal and noise) but subject to noisy moment representation (noisy conservation laws). Other examples include the class of regularization methods, which use the penalty function 2.13 for different levels of α (e.g., Donho et. al., 1992), or the class of models known as "quantified ME" (e.g., Skilling, 1989).[9] The solutions in all of these methods depend on the choice of α, and the moments' representation, and all are derived as in the traditional ME approach.

Finally, there are two common and important ingredients in all of these information-theoretic methods. First, they all are 100% efficient Information Processing Rules (IPR), in the sense defined by Zellner (1988). A 100% efficient IPR is one that satisfies the "information conservation principle" where the input information equals the output information. Thus, there is no loss of information in the inversion process. The two components of the input information are the data density (or likelihood function in the more traditional approach) and the prior distribution/s. The two output information components are the post-data distributions and the relevant partition functions (or

[9] A less well known class of ME-type methods that was developed for noisy data, known as the ME on the mean (and is related to the GME), is discussed in Gamboa and Gassiat (1997).

marginal probability distributions). Using Eq. 2.13, it is possible to show that even though each one of these information-theoretic methods is 100% efficient, the amount of input information may change according to the choice of $\boldsymbol{a}$ as well as the conservation laws (moment specification) used. These choices in turn affect the output information.

Second, the properties of all of these methods can be developed and compared via the theory of Large Deviations, LD. LD deals primarily with the convergence rates of stochastic systems and is based on the earlier work of Cramer (published in 1938) and Chernoff (1952). While the law of large numbers shows that certain probabilities converge to zero, LD theory is used to investigate the rate of convergence of these probabilities to zero. For example, with (exponentially) bounded random variables, the rate at which probabilities converge to zero rises exponentially as the size of the sample increases. These exponential decay rates are computed in terms of the entropy (or cross entropy) function and are different for each level of α in (2.13). In most (regular) cases, the second derivative of $I_{a+1}(\boldsymbol{p;q})$ evaluated at the mean is just the inverse of the variance (or the Fisher information matrix). Therefore, comparing different rate functions, which are specified in terms of entropy functions, gives information about asymptoptics. LD is often used in the areas of hypothesis testing (calculating the rate that the probability of making an error goes to zero), in fast simulation methods, and in comparing estimation methods on the basis of their convergence rate. For a detailed discussion, see the classic text by Ellis (1985) and the early work of Csiszar (1984). The non-iid case is developed in this volume by Kitamura and Stutzer.

2.4. Information, Entropy, Complexity and Non-Linearity

In addition to methods of estimation and inference within the above framework, there is a tight connection between IT, entropy, ME and analysis of complex and non-linear systems. In particular, versions of Eq. 2.13 are commonly used to investigate the linear and non-linear dependence among random variables. Quantities such as the Lyapounov exponents (measuring the non-linearity of a system and whether the system is chaotic or not), fractal and multi-fractal dimensions and correlation dimensions are just a few examples. All of these quantities describe the amount of information, or information decay, in a system and are used to investigate non-linear (dynamic) systems within parametric and non-parametric frameworks. For example, take the mutual information (defined as the expected information in an outcome of a random draw from $\boldsymbol{y}$ about an outcome of a random draw from $\boldsymbol{x}$) version of (2.13) for two discrete random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ of dimension N, and for α=1:

$$I_2^R(\boldsymbol{x},\boldsymbol{y}) \equiv H_2^R(\boldsymbol{y}) - [H_2^R(\boldsymbol{x},\boldsymbol{y}) - H_2^R(\boldsymbol{x})]. \qquad (2.14)$$

This measure equals zero if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent, and it equals *log (N)* if and only if $\boldsymbol{y}=f(\boldsymbol{x})$, where f can be any linear or nonlinear function. In general, this type of measure is used for any value of $\boldsymbol{a}$ where $\boldsymbol{a}$ is directly related to the system's (embedding) dimension, or where $\boldsymbol{a}$ is related to (multi) fractal dimension in a nonlinear-chaotic system. For more, see Soofi (1994), discussions by Maasoumi and Racine and by Ullah in this volume, as well as one of the many texts on non-linearities, entropy, and information.

3. Information and Entropy Econometrics and This Volume

The main areas discussed in this volume are information and information measures in general, the relationship of IT and Bayes, information-theoretic methods, and nonlinear and non-parametric methods. The issues of model and variable selection enter many of these discussions. However, as model and/or variable selection is an essential ingredient of all analysis done with random and incomplete data, and is commonly related to IT, it does appear directly or indirectly in almost all of the presented papers, and is not treated here as a specific area of research.

As all research within IEE is based on the notion of entropy and information, it seems natural to open this volume with Soofi and Retzer's comprehensive discussion of information measures. In addition to reviewing statistical information and information indices, they develop a unified framework for these indices. They start with the classical cross-entropy formulation and then discuss "optimal models" within the IT framework. The connection between the ME and ML is discussed as well and is extended to the GLS framework. Detailed examples and empirical applications are provided.

Zellner provides an illuminating discussion of information processing rules (IPR) and extends the notion of optimal such IPR rules that he developed in 1988 to dynamic optimal processing rules. These rules are 100% efficient, are derived by optimizing some information criterion, and, as is shown by Zellner, are naturally connected with Bayes' theorem.

With the objective of searching for a class of likelihood free methods during the last decade, there are an exponentially increasing number of papers connecting IT, estimation and inference. These papers are connected in a number of dimensions. First,

they all are based on optimizing a version (discrete or continuous) of Eq. 2.13. Second, the optimization is always with respect to the observed moments where these moments are viewed as pure or noisy. Bera and Bilias take us through a fascinating historical review of these types of estimation rules and provide an easy to follow trail of statistical developments within a unified setting. Their review takes us back to Pearson's Chi-squared goodness of fit test, the traditional method of moment, the ML, the connection between the ME and the c^2 and then proceed to the original work on EL and GMM as well as other information-theoretic methods. This review and perspective brings us up to current GMM type methods. From here, we progress via a number of papers presenting new results and new developments. These developments include (i) improved confidence intervals and tests for such models, (ii) new information-theoretic moment estimators and extensions, (iii) new interpretations of such models and (iv) applications and specialized cases.

Imbens and Spady develop improved confidence intervals for the GMM method that are based on empirical likelihood methods. These new intervals are constructed for the exactly identified and over identified moment conditions cases. They contrast the large and small sample behavior of these EL-based intervals with the standard GMM approach and demonstrate that for small samples the different methods provide significantly different intervals.

Ramalho and Smith examine and develop non-nested tests for competing moment condition models within the framework of the Generalized EL. They show how tests of non-nested hypotheses might be constructed from alternative estimators to the two-step GMM estimator, which is known to have poor properties in small or moderate-sized

samples. Ramalho and Smith address a substantial gap in the literature by providing adequate tests of non-nested hypotheses based on GMM.

Within the information-theoretic-GMM framework, van Akkeren, Judge and Mittelhammer develop a new estimation method for the case where the linear model may be ill-behaved or ill-conditioned (e.g., possibly weak instruments, multicollinearity, small samples) and where the orthogonality condition $E[X'\boldsymbol{e}] = \boldsymbol{0}$ does not hold. Their method falls within the class of EL estimators and makes use of the over-identified noisy (generalized) moment formulation (e.g., the GME) to assign the possible range for the parameters. This method may be viewed as a more robust version of the GME as it is entirely data dependent. This coordinate based formulation has attractive finite and asymptotic properties and can lead to post data densities along the lines of Zellner's BMOM.

Nevo abstracts away from the traditional assumption that the sample is a random draw from the population of interest. He assumes that the available data are a draw from some sampled population, rather than the "target population." Therefore, the empirical distribution is not a consistent estimate of the underlying data generation process. He proposes a way to re-weigh the sample. These weights are calculated by the inverse probability of the selection and thus make the sample a "representative" sample again. Within the information-theoretic methods, he develops a logistic selection model where the weights are constructed via Bayes rule.

Kitamura and Stutzer continue their earlier work on the unification of information-theoretic methods and the GMM. They develop the relationship between the cross-entropy projection and the linear projection to problems of estimation and

performance diagnostics for models in asset pricing. By using large deviation (LD) theory they are able to provide a "frequentist" interpretation to cross-entropy and thus to nicely connect the "traditional" GMM with IT. Further, their information theoretic LD approach is extended here from the iid to the non-iid case.

Within the information-theoretic framework, Kim shows that Hansen's optimal GMM estimator can be subsumed within the maximum limited information likelihood (LILH) estimation context. Through I–projecting[10] the LILH from a set of distributions consistent with the limited information, contained in the (pure) moment restrictions, this LILH approach allows the researcher to draw Bayesian inference, using less-than-fully parameterized specifications of the likelihood. The four main points Kim develops and discusses are that (i) a likelihood can be constructed on a subset of the parameters that are of interest using only limited moment information together with the I-projection functional, (ii) GMM is fully subsumed by the LILH approach, (iii) the results enable one to make a Bayesian inferences (including model selection), and (iv) this method may be applied to problems which GMM have addressed by, even in the absence of a parametric likelihood specification. Since the basic idea behind the LILH is related to Zellner's BMOM, Kim contrasts and compares the two.

Within a nested GME method, Golan and Perloff derive an axiomatic basis for a class of estimators. They start by nesting the GME within two more general (Tsallis and Renyi) entropy measures indexed by a single parameter α, (Eq. 2.13), which they call GME-α. Based on the earlier axiomatic derivations behind the ME, they show that the

[10] I-projection theory states that out of a set of probability measures satisfying the same moment conditions one chooses the probability measure that minimizes the entropy distance, or K-L distance, from the true probability measure. This is also known as the basic "cross entropy" formalism.

GME is the only method satisfying six natural axioms, while each of the more general GME-α estimators violates one of the basic axioms. The small sample behavior of the different estimators is demonstrated via sampling experiments.

Gregory, Lamarche and Smith continue the search for the small sample behavior of the information theoretic methods as compared with the iterated GMM. They build on the early work of Imbens et. al (1998) and Kitamura and Stutzer (1997) and provide sampling experiments evidence for both estimators. They make comparison in both the iid and non-iid worlds. They show that the information-theoretic method provides superior size-adjusted power. They conclude by applying these models to two macroeconomic time-series problems. The first problem assumes independence of moments over time while the second assumes dependency.

LaFrance, Beatty, Pope, and Agnew use IT and ME in order to infer the U.S. income distribution from data on quintile and top-five percentile income ranges as well as intra-quintile and top-five percentile mean incomes. These different ME income distributions are combined with data on the demand for different food items in order to estimate the overall incomplete system of demand from a long time-series of U.S. consumption data. Within the discrete and continuous ME methods, they compare different forms of representing the available information. The resulting ME distribution is either a piece-wise uniform density or a (smooth) continuous function.

Miller and Liu apply the ME principle to recover joint distributions from joint moments or marginal densities or both. They provide a nice review of current information approaches for solving this problem and then introduce a new minimum cross-entropy method for recovering joint densities from incomplete information.

Karantininis applies a version of the GME for analyzing a non-stationary transition probability matrix to examine adjustment in the Danish pork industry. Due to missing data problems, he uses the ME and GME methods to disaggregate these missing data in a first step analysis.

Maasoumi and Racine use a normalized entropy version of (2.13) to investigate possible non-linear relationships in stock returns. Their entropy metric is defined over the densities of stock returns that are estimated non-parametrically. They connect their measure in a very elegant way with the generalized entropy (2.13) and discuss the basic properties their measure satisfies. They use their information measure to predict excess returns of monthly stock returns.

As previously discussed, variations of the entropy and the K-L divergence measure are often used for both hypothesis tests and non-parametric methods. In his work, Ullah expands on both of these issues. First, he provides new results regarding the evaluation of the accuracy of approximations in general, such as evaluating the exact densities of some estimators and the relevant test statistics in finite sample econometrics. He compares the accuracy of alternative approximations. Ullah develops a general result for the non-iid random vector with at least four finite first moments where he shows that the K-L divergence measure is a special case of this result. In that way, he develops an asymptotic expansion of entropy and divergence functions. Second, Ullah applies the K-L measure toward non-parametric estimation and hypothesis testing such as non-parametric kernel estimation and the F-type non-parametric test statistics.

Ramsay and Ramsey develop a functional analysis of non-linear, time-series data. With the objective of identifying non-parametrically a set of differential equations that

capture the underlying dynamics of an economic production process, they are faced with a basic under-determined problem. Instead of directly employing some version of Eq. 2.13 subject to the observed information, they assume that even though the observed data are discrete the true underlying process is distributed smoothly and therefore can be characterized by a system of differential equations. With this basic assumption, they proceed to develop their new and innovative method, contrast it with the ME approach, and apply it to examine the dynamics of a monthly non-seasonally adjusted index of production, for the U.S. manufacturing, over a long period. In that way, they basically use more information than is commonly used with ME methods by incorporating a different penalty function and by introducing more structure. It would be interesting to compare their approach with the BMOM that also uses a continuous density, is fully characterized within the ME formalism, but replaces the (economic) information they use, with more information on the observed moments.

In order to estimate a labor supply function while permitting flexible preferences, van Soest, Das, and Gong develop a flexible non-parametric model of the utility function. They employ series expansions in hours and income to approximate the utility function. They use smooth simulated ML to evaluate their model for every given length of series expansion. Building on the likelihood ratio test and its relationship with the entropy-ratio statistic, they use the K-L information criterion to choose the optimal length of the series expansion. They choose the optimal length by generalizing Soofi's information index (or the generalized entropy measure) to also account for the unobserved wage heterogeneity in their model.

4. Conclusion

The current work in Information and Entropy Econometrics (IEE) can fill volumes upon volumes of journals. Even though we could not capture all of the interesting new results here, I believe this volume contains a representative sample of the work on Information Theory and Maximum Entropy (ME) within IEE. This collection of work includes both historical and logical perspectives as well as current state of the art research in the area. Where do we go from here, only the future knows, but it seems safe to predict that information theory, entropy and ME will continue to play an important role in the development of econometric thoughts, theory and applications.

Acknowledgements

I am grateful to the authors for not only making the deadlines, but mostly for providing excellent contributions that help us in better understanding and connecting the more traditional econometrics, IT, and ME. I must acknowledge the countless discussions, debates and joint work throughout the years that I have enjoyed with George Judge. It is due to these discussions and joint work that I was able to process some of the information within IT and ME and to be able to connect it with the more traditional econometrics. I am very thankful to Arnold Zellner for his support and encouragement throughout the years and for his help that was essential in putting this volume together. I am especially thankful for more than forty reviewers, referees, and associate editors whose specific comments and suggestions improved the content of each one of the papers and the general content of this volume as a whole. I am also grateful to Dose Volker, Elise Golan, Ehsan Soofi, Essie Maasoumi, Ron Mittelhammer, Jeff Perloff, George Judge,

Arnold Zellner, Anil Bera, Doug Miller, Yannis Bilias, Jing Qin and Constantino Tsallis for their helpful comments and feedback on earlier versions of this essay.

References

Chernoff, H., 1952. A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. Annals of Math. Stat., 23, 493-507.

Cover, T. M. and J. A. Thomas, 1991. Elements of Information Theory. John Wiley & Sons, New York.

Cressie, N and T. R. C. Read, 1984. Multinomial goodness-of-fit tests. J. Royal Stat. Soc. B, 46, 440-464.

Csiszar, I., 1984. Sanov property, generalized I-projection and a conditional limit theorem. Annals of Probability, 12, 768-793.

Davis, H. T., 1941. The Theory of Econometrics. The Principia Press, Indiana.

Donoho, D. L., I. M. Johnstone, J. C. Hoch, and A. S. Stern, 1992. Maximum entropy and the nearly black object. Journal of the Royal Statistical Society, Ser. B, 54, 41-81.

Ellis, R. S., 1985. Entropy, Large Deviations, and Statistical Mechanics. Springer-Verlag, New York.

Gamboa, F. and Gassiat, E. 1997. Baysian methods and maximum entropy for ill-posed inverse problems. Annals of Statistics. 25(1) ,328-350

Goldstine H. H., 1980. A History of the Calculus of Variation from 17th Through the 19th Century. Springer-Verlag, New York.

Imbens, G.W., Johnson, P. and R.H. Spady, 1998. Information-Theoretic Approaches to Inference in Moment Condition Models. Econometrica 66, 333-357.

Jaynes, E.T., 1957a. Information theory and statistical mechanics. Physics Review, 106, 620-630.

Jaynes, E.T., 1957b. Information theory and statistical mechanics II. Physics Review, 108, 171-190.

Kitamura, Y. and M. Stutzer, 1997. An information-theoretic alternative to generalized method of moment estimation. Econometrica 66 4, 861-874.

Kullback, S., 1954. Certain inequalities in information theory and the Cramer-Rao inequality. The Annals of Math. Stat. 25, 745-751.

Kullback, S., 1959. Information Theory and Statistics. John Wiley & Sons, New York.

Kullback, S., and R. A. Leibler 1951. On information and sufficiency. The Annals of Math. Stat. 22, 79-86.

Lindley, D. V., 1956. On a measure of the information provided by an experiment. The Annals of Math. Stat. 27, 986-1005.

Maasoumi E., 1993. A compendium to information theory in economics and econometrics. Econometric Reviews 12, 137-181.

Renyi, A., 1961, On measures of information and entropy. Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability, 1960, vol. I, 547.

Renyi, A., 1970. Probability Theory. North-Holland, Amsterdam.

Sagan H., 1993. The Calculus of Variation. McGraw-Hill, New York.

Shannon, C. E, 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379-423.

Skilling J., 1989. Classic Maximum Entropy. In: Skilling, J. (Ed.) Maximum Entropy and Bayesian Methods in Science and Engineering. Kluwer Academic, pp. 45-52.

Soofi, E. S., 1994. Capturing the intangible concept of information, J. of the American Statistical Association, 89, 1243-1254.

Tsallis, C., 1988. Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys., 52, 479-487.

Zellner, A., 1988. Optimal information processing and Bayes theorem. American Statistician, 42, 278-284.

Zellner, A., 1991. Bayesian methods and entropy in economics and econometrics. In Grandy, W.T., Jr. and Schick, L. H. (Eds.), Maximum Entropy and Bayesian Methods, Kluwer, Amsterdam, 17-31.

Zellner, A., 1997. The Bayesian method of moments (BMOM): Theory and applications, in T. Fomby and R. Hill, eds., Advances in Econometrics 12, 85-105.