ELSEVIER

Journal of Economic Dynamics and Control
20 (1996) 819–824

JOURNAL OF
Economic
Dynamics
& Control

Elements of information theory
A book review

Shu-Heng Chen

National Chengchi University, Taipei, Taiwan 11623, Republic of China

(Received August 1995)

Tomas M. Cover and Joy A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, NY, 1991)

The 'information' used among economists appear in primarily three different contexts: the *resource allocation of information*, such as Stigler (1964), the *measure of information* and its applications, such as Theil (1967)[1], and the *transmission of information*, such as O'Neill (1987). It is in the last two arenas where economists may feel motivated to read Cover and Thomas. Since most economists are more familiar with Theil (1967), I would like to review Cover and Thomas' book by comparing it with Theil's. The comparison is centered upon the following two fundamental questions:

1. What is information?
2. Why is information theory significant?

What is information? The 'information' used by Theil and his followers were largely restricted to the information theory before 1960, i.e., the foundation laid by Shannon, Wiener, Weaver, Kullback, Leibler, Fano, and Jayne, which is also called *Shannon information theory*. This theory uses probability theory as its theoretic foundation. For example, it constructs *entropy* from the probability mass function or density function of a random variable. Cover and Thomas also take this approach. In Chapters 1 and 2, *entropy*, *joint entropy*, *relative entropy*,

[1] For a survey article on this direction, please see Massoumi (1990).

SSDI 0165188995008 77 X

as well as *mutual information* are all defined from the probability mass function. Moreover, in Chapter 9, they develop the concept of *differential entropy* for the continuous random variable.

However, they do not stop there. Cover and Thomas also incorporated in their book the concepts of *algorithmic information theory*. Developed independently by Solomonof, Kolmogorov, and Chaitin, algorithmic information theory is one of the most important developments since 1964. This approach re-examines the foundation for information theory. Instead of treating probability theory as the foundation of information theory, it considers that information theory must precede probability theory.[2] Nevertheless, this approach had been largely neglected by most textbooks in Theil's tradition, and had hence been neglected by most textbooks in statistics and econometrics. Cover and Thomas have no intention to bypass it. A whole chapter (Chapter 7) is devoted to *Kolmogorov complexity*, one of the most important concepts in algorithmic information theory.

Chapter 7 is written in such a compact way that it is easy for beginners to have a quick grasp of the essential ideas in algorithmic information theory without being troubled by technical details. Roughly speaking, Kolmogorov complexity is a measure for the descriptive complexity of an object. We say an algorithm can describe an object if the object can be the output of the *universal Turing machine* when the algorithm is fed to the machine. Without loss of generality, the algorithm can be coded by binary digits. Thus, each algorithm is nothing but a string and the length of a string is simply the number of the binary digits in that string. The Kolmogorov complexity of an object is defined to be the shortest string that can describe the object.

If we use Kolmogorov complexity rather than entropy as the definition or measure for information, then it is clear that probability is not indispensable for the scientific conceptualization of information. But, without probability, can we still talk about 'random'? The authors answer this question by introducing the concept *algorithmic randomness* (Section 7.5). Roughly, an object represented in terms of binary digits is said to be random if its Kolmogorov complexity is equal to or larger than its length. Therefore, algorithmic information theory captures the meaning of randomness in a most intuitive way, i.e. the absence of periodicity.

After introducing the two most important concepts in algorithmic information theory, i.e., Kolmogorov complexity and algorithmic randomness, the authors start to inquire into the relationship between the foundation of algorithmic information theory and that of Shannon information theory. Two main

[2] As Kolmogorov (1983, p. 31) stated: 'From the general considerations that have been briefly developed it is not clear why information theory should be based so essentially on probability theory, as the majority of text-books would have it. It is my task to show that this dependence on previously created probability theory is not, in fact, inevitable.'

results are stated as follows. Firstly, in Section 7.3, they show that, when we apply Kolmogorov complexity to an object within a probability framework, e.g., an ensemble from a well-defined stochastic process, the expected value of the Kolmogorov complexity of a random sequence turns out to be close to Shannon entropy (Theorem 7.3.1). Secondly, in Section 7.5, they prove the strong law of large numbers for incompressible sequences, which means that incompressible sequences look random in the sense that the proportion of 0's and 1's are almost equal. A more general version of this result is that the algorithmic test for randomness is the ultimate test.

As the authors said in their preface, Chapter 7 has no counterpart in other information theory texts. To organize such an abstruse topic so well in a single chapter is what makes this book unique.

Having given the two definitions of information, we then come to the next issue, the significance of information theory. The beauty of Cover and Thomas is its way of treating information theory not just as a subset of communication theory, but as a powerful language to be used in different branches of science such as physics, computer science, probability, statistics, and economics. I shall elaborate on this point from an economist's perspective.

One of the most important contributions of information theory is its application in finance. Chapter 6, for example, illustrates the implications of information theory for gambling (horse races). There is strong duality between the growth rate of the gambler's wealth and the entropy rate of the horse race. The lower the entropy rate, the higher the growth rate.[3] Moreover, a good gambler is also a good data compressor in that the gambler's bets can be considered to be his estimate of the probability distribution of the data. For the gambler, his growth rate of wealth is influenced by his estimate of the true distribution.[4] The better the estimate, the higher the growth rate. For the data compressor who is using arithmetic coding, how well he can compress his data depends on the distribution used to compress the data. If he is using the true distribution, then by Shannon coding theorem, the entropy rate is the low bound for his data compression. If he does not know the true distribution, the better he estimates, the more he can compress.

Apart from finance, information theory has also made significant contributions to econometrics. As a teacher of the first course in graduate econometrics, I usually begin by reviewing statistical decision theory. Within this framework, the conditional expectation is an optimal decision in the sense that it minimizes expected risk if the loss function is quadratic. Furthermore, under *joint Gaussian*

[3] To have this result, the gambler is required to use the log-optimal gambling strategy associated with 100% reinvestment in uniform-odds horse races.

[4] This can be measured by relative entropy, see p. 128.

assumption, conditional expectations are equivalent to linear regression models.[5] This provides a justification for the use of linear regression models at the beginning of most econometric texts. However, bright students will challenge you by asking: 'Then what is the justification for using joint Gaussian assumption?' A more fundamental question would be how we choose models. It is from this question that I realize the indispensible role of information theory in the foundations of econometrics. In fact, up to the present, the only solid foundation on which model selection principles have been built is information theory.

To prove my point, I would like to bring up two principles used in econometrics which have root in information theory. One is the *maximum entropy principle*, which is associated with *Shannon information theory*. The other is the *minimum description length (MDL) principle*, which derives from *algorithmic information theory*.[6] As a textbook, Cover and Thomas pave for the readers a pretty smooth trail to the understanding of how these two principles have root in information theory.

In Chapter 11, joint Gaussian assumption N(0, K) is shown to be the *maximum entropy distribution* given that the multivariate random vector X has zero mean and variance–covariance matrix $E(XX') = K$.[7] Therefore, the joint Gaussian distribution is *minimally prejudiced* in the sense of Jaynes (1957) which is '*maximally non-committal with regard to missing information*'.[8] Moreover, by Burg's theorem introduced in Section 11.6, the familiar AR(p) Gaussian processes in time series analysis can also be derived by the maximum entropy principle given appropriate constraints.

While the ME principle is well-known, the MDL principle (or stochastic complexity) is still new for most econometricians.[9] The reason for this is simple. The foundation for algorithmic information theory is algorithmic complexity in computation theory which is seldom a part of training for economists. Therefore, econometricians tend to have some difficulty in understanding this principle. For those economists who want to be motivated to study the MDL

[5] This approach is similar to Amemiya (1985) though he has a little different order.

[6] One may wonder why the celebrated maximum likelihood (ML) principle is not singled out. There are two reasons for this. First, the MDL principle can be equally justified as the 'global ML principle' (see Rissanen, 1989, p. 6). Second, it is well-known that the ML principle is a guidance for parameter estimation under chosen models and cannot be considered a general principle for model selection.

[7] This is also shown in Chapter 9, Theorem 9.6.5.

[8] See Jaynes (1957, p. 620).

[9] Fortunately, in the recent literature of bounded rationality in economics, it has caught the attention of a number of economists. For example, Leijonhufvud (1993, p. 6) was aware of the importance of this principle: 'We should not look at Rissanen's work as 'only' providing a new statistical foundation to the econometricians; it also offers theorists a way to populate their models with agents that learn by *induction*.'

principle, Sections 7.6–7.11 provide some useful background knowledge. The way the authors present these sections is rather entertaining. I am particularly amused by their illustration of the concept of *universal probability*. The authors invite the readers to imagine a monkey sitting at a computer keyboard and typing the keys at random. The inquiry about the chance that all the works of Shakespeare are typed out immediately gives us an idea of what universal probability is about. After giving a definition of universal probability in Section 7.6, the authors prove an equivalence between Kolmogorov complexity and universal probability in Section 7.11. The content of universal probability is shown to be further enriched by its contribution to the understanding of Occam's Razor, Chaitin's Ω, and Fermant's last theorem. Moreover, the relations between algorithmic information theory and statistics are illuminated by the fact that such important statistical concepts as the likelihood ratio test (Section 7.6), LaPlace's problem (Section 7.10), and sufficient statistics (Section 7.12) can be redeveloped from universal probability.

In terms of the relationship between the MDL principal and algorithmic information theory, the authors claim that 'Rissanen's minimum description length (MDL) principle is very close in spirit to the Kolmogorov sufficient statistic' (p. 182). This seems to be an oversimplifying statement. In fact, the major issue that concerns Rissanen is the concrete application of algorithmic information theory to statistical inference.[10] Therefore, while they are 'close in spirit', their difference also deserves attention. As Rissanen (1992, pp. 2–3) states: 'The stochastic complexity differs from the algorithmic complexity mainly in the nature of the models and the model classes chosen, in which computability theory plays no role.' Cover and Tomas' treatment on information theory and statistics would be complete if Rissanen's work were covered in a little more detail.

The rest of the book is mainly related to communication theory where Shannon's first, second, and third theorem are established. While it is not difficult to realize that *channel* and *market* are interchangeable in terms of information transmission, the relevance of communication theory to economics remains to be seen. To my best knowledge, O'Neill (1987) appeared to be the first application of *Shannon's coding theorem* in the area of economic modelling.[11] He showed that Shannon's coding theorem allows one to precisely define the sense in which information is transmitted by an economic market (channel). With the help of *rate distortion theory*, he was able to show that the

[10] This can be clearly seen at the very beginning of Rissanen (1989, p. i) where he writes: 'However, the theory gives no guidance to the practical construction of programs, let alone of the shortest one, which in fact turns out to be non-computable. Accordingly, the theory has had little or no direct impact on statistical problems.'

[11] I am grateful to Professor Velupillai for bringing my attention to the existence of such research.

Grossman–Stiglitz equilibrium requires signaling rate in excess of the channel capacity. To gain familiarity with this kind of research, the reader will find Chapters 8, 10, and 13 of the book helpful.

In sum, given these characteristics, Cover and Thomas provide a much more general interdisciplinary framework for information theory than most of the other texts. It makes readers more curious about the significance of information theory. In terms of a textbook, it is certainly a great success.

References

Amemiya, T., 1985, Advanced econometrics (Harvard University Press, Cambridge, MA).

Jaynes, E.T., 1957, Information theory and statistical mechanics: I, II, Physical Review 106, 620–630 and Physical Review 108, 171–190.

Kolmogorov, A.N., 1983, Combinatorial foundations of information theory and the calculus of probabilities, Russian Mathematical Surveys 38 (4), 29–40.

Leijonhufvud, A., 1993, Towards a not-too-rational macroeconomics, Southern Economic Journal 60, 1–13.

Massoumi, E., 1990, Information theory, in: J. Eatwell, M. Milgate, and P. Newman, eds., Econometrics (Norton, New York, NY).

O'Neill, W.D., 1987, An application of Shannon's coding theorem to information transmission in economic markets, Information Sciences 41, 171–185.

Rissanen, 1989, Stochastic complexity in statistical inquiry (World Scientific, Singapore).

Rissanen, 1992, Stochastic complexity, information, and learning, Paper presented in the UCLA computable economics workshop (University of California, Los Angeles, CA).

Stigler, G.J., 1964, The economics of information, Journal of Political Economy 69, 213–225.

Theil, H., 1967, Economics and information theory (North-Holland, Amsterdam).