# Capturing the Intangible Concept of Information

Ehsan S. Soofi*

The purpose of this article is to discuss the intricacies of quantifying information in some statistical problems. The aim is to develop a general appreciation for the meanings of information functions rather than their mathematical use. This theme integrates fundamental aspects of the contributions of Kullback, Lindley, and Jaynes and bridges chaos to probability modeling. A synopsis of information-theoretic statistics is presented in the form of a pyramid with Shannon at the vertex and a triangular base that signifies three distinct variants of quantifying information: discrimination information (Kullback), mutual information (Lindley), and maximum entropy information (Jaynes). Examples of capturing information by the maximum entropy (ME) method are discussed. It is shown that the ME approach produces a general class of logit models capable of capturing various forms of sample and nonsample information. Diagnostics for quantifying information captured by the ME logit models are given, and decomposition of information into orthogonal components is presented. Basic geometry is used to display information graphically in a simple example. An overview of quantifying information in chaotic systems is presented, and a discrimination information diagnostic for studying chaotic data is introduced. Finally, some brief comments about future research are given.

KEY WORDS: Chaos; Entropy; Kullback–Leibler; Logit; Principal components.

## 1. NOTIONS OF INFORMATION

In Webster's Third New International Dictionary, the definitions of "information" include "the communication or reception of knowledge and intelligence," "knowledge communicated by others and/or obtained from investigation, study, or instruction," "facts and figures ready for communication or use as distinguished from those incorporated in a formally organized branch of knowledge, DATA," "the process by which the form of an object of knowledge is impressed upon the apprehending mind so as to bring about the state of knowing," and "a numerical quantity that measures the uncertainty in outcome of an experiment to be performed." The last definition refers to the information entropy. Thus the notions of information consist of a spectrum ranging from semantic to technical. In the semantic context, the term information is used in an intuitive sense. It does not refer to a well-defined numerical quantity that can be used for measuring the extent of uncertainty differentials due to changes in the states of nature. In the technical sense, information is referred to as a well-defined function that quantifies the extent of uncertainty differentials. In statistics and scientific literature involving statistical analysis, the term information is often used ambiguously. Sometimes information is used in a semantic sense while the context requires a very precise technical notion.

Statisticians have long endeavored to develop a precise notion of information. The earliest and most well-known technical definition of information was given by Fisher (1921), who, in the context of parametric estimation, defined the inverse of the variance of the sampling distribution of an estimate as the measure of *relevant* information provided by data about an unknown parameter. This has led to the use of inverse of variance as a measure of information contained in a distribution about the outcome of a random draw from that distribution. Fisher's information is a very special case of a more general definition later developed in statistical information theory.

Information theory has attracted the attention of researchers from various disciplines who are intrigued by "the apparent impossibility of capturing the intangible concept of information" (Cover and Thomas 1991, p. viii). Statisticians have been pivotal in the development of information theory and have shown that it provides a framework for dealing with a wide variety of statistical problems in a unified manner (Brockett 1991; Kullback 1959). But in general, the emphasis of information-theoretic statistics has been on providing alternative formulations to the traditional statistical analyses, and not much attention has been given to the elegance of quantifying information in specific problems. Many statisticians are familiar or are aware of information-theoretic approaches to discrimination between alternative models, testing goodness of fit, and developing prior distributions for Bayesian analysis. The purpose of this article is to highlight the intricacies of quantifying information in some statistical problems. I will discuss the problem of capturing information in two not so closely related problems: probabilistic-choice modeling and chaos. I will elaborate and extend information diagnostics developed for these problems in recent years.

The article is organized as follows. Section 2 presents a synopsis of the history of information-theoretic statistics with emphasis on three basic methods of quantifying information. Section 3 elaborates on the maximum entropy method by providing illustrative examples and new information diagnostics for choice modeling. Section 4 discusses capturing information in chaos. Finally, Section 5 points out some directions for future research.

* Ehsan S. Soofi is Professor, School of Business Administration, University of Wisconsin, Milwaukee, WI 53201. This article was developed based on the author's presentation at the 43rd Meeting of NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics at MIT, Cambridge, MA, November 1991; his presentation in Honor of E. T. Jaynes at the 44th Meeting of NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics at Washington University, St. Louis, MO, April 1992; and his discussion of the *JASA* Theory & Methods Invited Paper ("Likelihood and Bayesian Prediction of Chaotic Systems" by Mark Berliner at the 1992 Joint Statistical Meetings, Boston, August 1992). The author thanks the participants in these sessions, particularly Arnold Zellner, Jim Press, Tom Leonard, Wes Johnson, Asit Basu, Refik Soyer, Nozer Singpurwalla, and Ed Wegman, who expressed enthusiasm about this topic and encouraged the preparation of this article for the community at large. The author also thanks Nader Ebrahimi, Tim Haas, and two anonymous referees for their helpful comments on previous drafts of this article, and Joe Retzer for preparing the graphs.

Journal of the American Statistical Association
December 1994, Vol. 89, No. 428, Theory and Methods

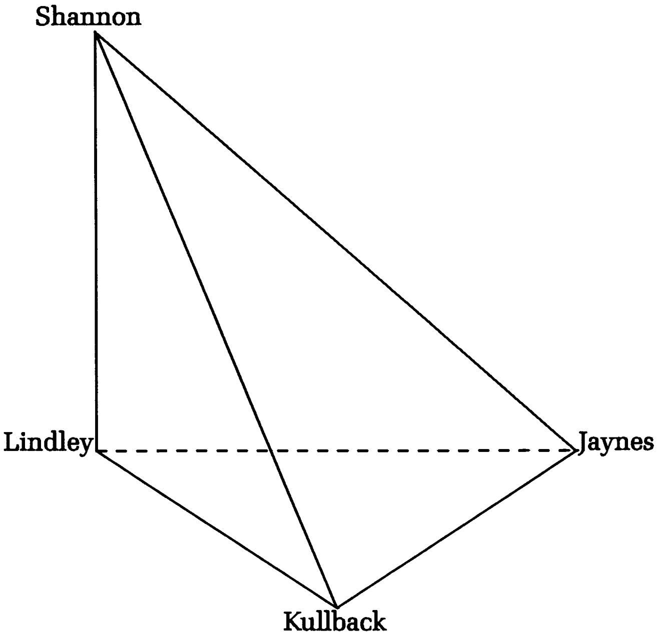

*Figure 1. Pyramid of History of Information-Theoretic Statistics.*

## 2. PYRAMID OF INFORMATION-THEORETIC STATISTICS

A synopsis of information-theoretic statistics may be represented by the Shannon–Kullback–Lindley–Jaynes (SKLJ) pyramid shown in Figure 1. At the vertex is Shannon, who developed the notion of information *entropy* in communication engineering. The base of the pyramid displays the immediate extensions of Shannon's work for statistical applications. Kullback pioneered the extension, and his work became dominant in statistical information theory. Lindley gave the most direct translation of Shannon's work in statistics. Jaynes introduced the maximum entropy principle of inference with which many statisticians have some familiarity but for which the statistics community as a whole has not yet developed sufficient appreciation. We will see that the discrimination information, the mutual information, and the ME formulation offer three distinct variants of quantification of information for various statistical analyses. The lateral faces of the pyramid are the SKJ minimum discrimination information plane, the SLK mutual information plane, and the SLJ Bayesian information theory plane. Most entropy-based statistical work may be located at one of the lateral faces. Akaike's information diagnostic and some other information-theoretic work may be placed in the interior of the pyramid. A number of generalizations of Shannon's entropy and some other information measures located at the exterior of this pyramid have been introduced in the literature. But because of the simplicity and the additive property of Shannon's entropy and its derivatives, they have become prominent information measures in statistics and many other fields.

### 2.1 Shannon

Shannon (1948) developed information entropy for quantifying the expected uncertainty associated with an outcome from a set of symbols $\{x_j, j = 1, \ldots, J\}$ that are received directly from a source $X$ according to a probability distribution $\pi(x)$. He showed that the unique (up to a constant) function of the probabilities that satisfies a set of intuitively appealing axioms is

$$H(X) = H(\pi_X) = -\sum_X \pi(x)\log \pi(x) \geq 0. \qquad (1)$$

The name entropy was chosen because of the similarity of (1) with the thermodynamic entropy expression. The notations $H(X)$, $H(\pi_X)$, $H(\pi)$, and $H(X|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$ being parameters of $\pi$, will be used interchangeably as suits the context. Entropy is a smooth concave function of the probabilities with a maximum of $\log J$ at $(1/J, \ldots, 1/J)$. Figure 2 displays the scaled entropy $H(\pi)$ for $J = 3$. The entropy of a distribution with infinite support may be infinite.

For a distribution with a continuous density, differential entropy is defined by

$$H(X) = -\int \pi(x)\log \pi(x)\, dx.$$

The differential entropy shares some but not all the properties of the discrete entropy. In particular, for continuous distributions, $H(X)$ is not scale invariant, because $H(cX) = \log |c| + H(X)$, but it is translation invariant, because $H(c + X) = H(X)$. The differential entropy may be negative and infinite. Boundedness of $\pi(x)$ implies $H(X) > -\infty$, and $\text{var}(X) < \infty$ implies $H(X) < \infty$ (Ash 1965, p. 237). But, for a distribution with a finite entropy, the variance may not exist. Figure 3 shows the entropy of a Pareto family with density $\pi(x|\alpha, \beta) = \alpha\beta^{-1}(x/\beta)^{-\alpha-1}$, $x > \beta$, $\alpha, \beta > 0$ as function of the parameters. Note that $H(X|\alpha, \beta) = \log(\beta) + 1/\alpha - \log(\alpha) + 1$ is finite over the entire parameter space, but $\text{var}(X|\alpha, \beta) = \beta^2\alpha(\alpha - 1)^{-2}(\alpha - 2)^{-2}$ is not defined when $\alpha < 2$. Consequently, the variance cannot be used for comparing uncertainties associated with two Pareto distributions when $\alpha < 2$ for one or both distributions. But comparison of uncertainties associated with any two Pareto dis-

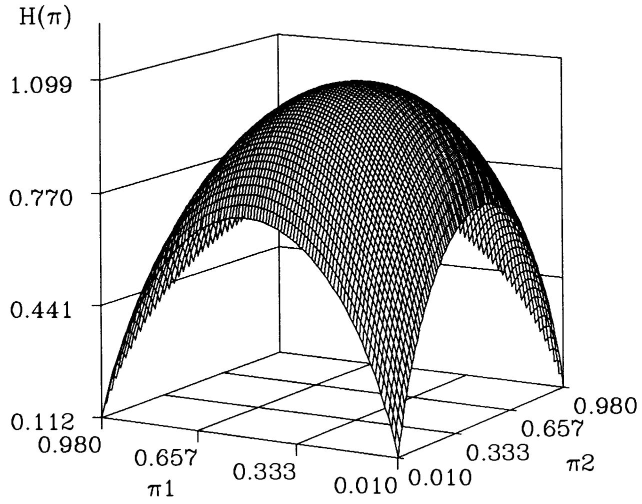

*Figure 2. The Entropy of the Distribution $\pi = (\pi_1, \pi_2, \pi_3)$ as a Function of the Probabilities.*

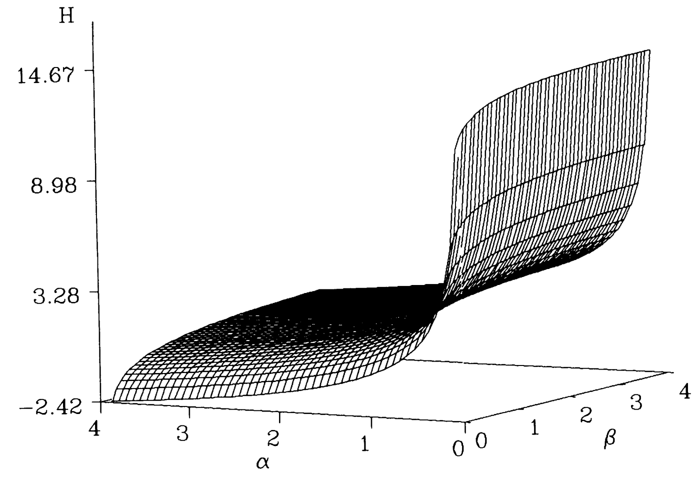

*Figure 3. The Entropy of the Pareto Distribution as a Function of the Parameters $\alpha$, $\beta > 0$.*

tributions are possible based on the entropy difference $\delta H[\pi_1(x|\alpha_1, \beta_1), \pi_2(x|\alpha_2, \beta_2)] = H[\pi_1(x|\alpha_1, \beta_1)] - H[\pi_2(x|\alpha_2, \beta_2)]$. When $\delta > 0$, $\pi_1$ is less informative than $\pi_2$, irrespective of the signs of the individual entropies involved.

The entropy summarizes the uncertainty associated with a distribution $\pi(x)$ in terms of concentration of probabilities. Thus it provides information about the predictability of an outcome of $X$. When the distribution depends on a set of parameters $\boldsymbol{\theta}$, $H(X|\boldsymbol{\theta})$ may not be interpreted as a measure of information about $\boldsymbol{\theta}$. But when $X$ is a suitable estimate for $\theta$, one may interpret information about $X$ as the information about $\theta$ (see Ebrahimi and Soofi 1990 for an example).

A noiseless communication channel is characterized by a set of input symbols $y$, a set of output symbols $x$, and the conditional distribution $\pi(y|x)$. If $y$ is transmitted from the source and $x$ is received, then the conditional entropy $H(Y|X = x)$ is computed by using $\pi(y|x)$ in (1). The expected uncertainty about a random draw from $Y$ given a random draw from $X$ is measured by $H(Y|X) = E_x[H(Y|X = x)]$.

The change of uncertainty about $Y$ due to observing $x$ is measured by $\delta H[\pi(y), \pi(y|x)] = H[\pi(y)] - H[\pi(y|x)]$. This quantity measures the amount of information contained in a particular $x$ about an outcome of a random draw from $Y$. In general, $\delta H[\pi(x), \pi(x|y)]$ may be positive or negative. The expected information in an outcome of a random draw from $X$ about an outcome of a random draw from $Y$ is given by the *mutual information*

$$\vartheta(Y|X) = E_x\{\delta H[\pi(y), \pi(y|X = x)]\}$$
$$= \sum_y \sum_x \pi(x, y) \log \frac{\pi(x, y)}{\pi(x)\pi(y)} \geq 0. \qquad (2)$$

The supremum of $\vartheta(Y|X)$ over the set of input distributions is defined as the capacity of the channel.

Note that $\vartheta(Y|X) = \vartheta(X|Y)$ and $\vartheta(X|Y) = 0$ if and only if $X$ and $Y$ are independent. The mutual information, therefore, quantifies information about the predictability of one variable given the other. The mutual information function is related to the joint, conditional, and marginal entropies as

$$\vartheta(Y|X) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$
$$= H(X) + H(Y) - H(X, Y).$$

The basic properties of $\vartheta(X|Y)$ for continuous and discrete distributions are the same.

Statisticians have used the mutual information at least in two contexts: measurement of dependency between variables (see, for example, Bozdogan 1990 and Joe 1989) and Bayesian comparison of experiments due to Lindley (1956), to be discussed in Section 2.3.

## 2.2 Kullback

Kullback and Leibler (1951) generalized (1) and (2) to an abstract level by defining

$$D(\pi_1 : \pi_2) = \int \log \frac{\pi_1(x)}{\pi_2(x)} \, d\Pi_1(x) \geq 0, \qquad (3)$$

where $\pi_i$ is the probability density (mass) function of $\Pi_i$, $i = 1, 2$. $D(\pi_1 : \pi_2)$ is defined for $\pi_1(x) = 0$ whenever $\pi_2(x) = 0$.

The Kullback–Leibler function (3), also known as information number, divergence, and "distance," is the entropy of $\pi_1$ relative to $\pi_2$. The relative entropy (i.e., cross-entropy) is convex function of the pair $(\pi_1, \pi_2)$, but it is not symmetric in $(\pi_1, \pi_2)$; $\pi_2$ is the *reference* distribution. Jeffreys (1946) considered a symmetric version of this function as a measure of divergence between two distributions with densities $\pi_1$ and $\pi_2$, $J(\pi_1 : \pi_2) = D(\pi_1 : \pi_2) + D(\pi_2 : \pi_1)$. Note that $D(\pi_1 : \pi_2)$ is invariant under one-to-one transformation of $x$. The properties of $D(\pi_1 : \pi_2)$ for continuous and discrete distributions are alike. Figure 4 displays normalized $D(\pi_1 : \pi_2)$ for discrete distributions over three outcomes when $\pi_2 = \zeta = (.1, .6, .3)$.

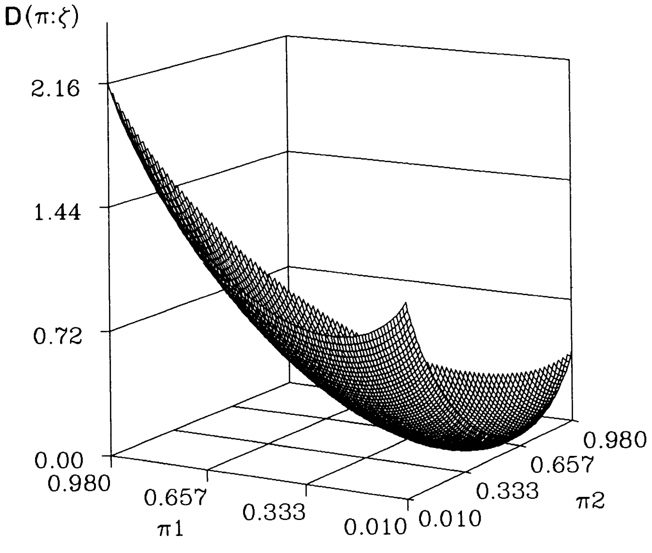

*Figure 4. The Discrimination Information Between the Distributions $\pi = (\pi_1, \pi_2, \pi_3)$ and $\zeta = (.1, .6, .3)$ as a Function of the Probabilities.*

The relative entropy is a measure of discrepancy between two distributions, and its interpretation as an information quantity depends on the distributions involved in the context of the problem under consideration. In a given problem, it may be used as a measure of information loss or gain due to change from $\pi_2$ to $\pi_1$. The mutual information (2) is also relative entropy $D[\pi(x, y) : \pi(x)\pi(y)]$, which is interpreted as information in one variable for predicting the other. Kullback's interpretation of (3) is, almost exclusively, the *mean information in x for discrimination between $H_1$ and $H_2$* in the following sense. Bayes's theorem gives

$$\log \frac{\pi_1(x)}{\pi_2(x)} = \log \frac{P(H_1 | x)}{P(H_2 | x)} - \log \frac{P(H_1)}{P(H_2)}, \qquad (4)$$

where under the hypothesis $H_i$, $i = 1, 2$, $\pi(x) = \pi_i(x)$; $P(H_i)$ and $P(H_i | x)$ are the prior and the posterior probabilities of $H$. Good (1950) referred to $\log[\pi_1(x)/\pi_2(x)]$ as the *weight of evidence.* Kullback and Leibler (1951) based on (4) interpreted $\log[\pi_1(x)/\pi_2(x)]$ as the *information in x for discrimination between $H_1$ and $H_2$*; thus $D(\pi_1 : \pi_2)$ is interpreted as the expected information in an observation from $X$ for discrimination between the two hypotheses.

Kullback and Leibler (1951) explicated information sufficiency in terms of (3). Suppose that $X$ is a sample from the family of distributions with generalized densities $\{\pi_i(x | \theta), i = 1, 2\}$, and let $Y = T(X)$ be a statistic. If $T$ is a measurable function and $q_i(y) = \pi_i[T^{-1}(y)]$, then $D[\pi_1(x) : \pi_2(x)] \geq D[q_1(y) : q_2(y)]$, with equality if and only if $Y$ is a sufficient statistic for $\theta$. Thus information in a sample for discrimination "cannot be increased by any statistical operations and is invariant (not decreased) if and only if sufficient statistics are employed" (Kullback and Leibler 1951).

Fisher's notion of information within a second-order approximation is the discrimination information between two distributions that belong to the same parametric family and differ infinitesimally over a parameter space; that is, in (3), let $\pi_1(x) = \pi(x, \theta)$ and $\pi_2(x) = \pi(x, \theta + \Delta\theta)$.

Kullback (1954) introduced the *minimum discrimination information* (MDI) principle of information efficiency for statistical analysis in terms of (3). Information efficiency is a generalization of efficiency in terms of the Cramer–Rao lower bound. Kullback (1959) thoroughly discussed the foundations of information-theoretic approach to statistics and showed that MDI approach provides a unified treatment of many statistical methods, such as classical hypothesis testing and estimation, analysis of contingency tables, regression analysis, and traditional multivariate techniques.

The Kullback–Leibler discrimination function is widely used in statistics (see, for example, Akaike 1973; Gokhale and Kullback 1978; Johnson 1987; Johnson and Geisser 1983; Kullback 1959; McCulloch 1989; Sawa 1978; Soofi and Gokhale 1991). But (3) is often used just as a measure of divergence between two probability distributions rather than as a meaningful information quantity in the context of the problem being discussed.

## 2.3 Lindley

Lindley (1956) gave the most direct translation of Shannon's communication systems into statistics. In statistics, a parameter $\theta$ is the signal transmitted from an unobservable source (i.e., the parameter space) $\Theta$ according to the prior density $\pi(\theta)$. The data $x$ is viewed as the signal received with a sampling distribution $\pi(x | \theta)$. The information differential $\delta H[\pi(\theta), \pi(\theta | x)]$ quantifies the amount of information in an observed sample $x$ about $\theta$. If this quantity is positive, then the sample is informative; if it is negative, then the sample increases the uncertainty, so the result is a "surprise." But before observing the data, the expected information in a random draw from $X$ about the parameter is given by the mutual information $\vartheta(\Theta | X) \geq 0$. Thus, on average, the samples are informative about the parameter. Lindley proposed the use of this information function for comparison of experiments. Because the objective of the Bayesian analysis is to learn about a parameter $\theta$ from $x$, the interpretation of $\vartheta(\Theta | X)$ is in the direction of $X$ to $\theta$ rather than mutual information between $\Theta$ and $X$. Bernardo (1979a) provided an interpretation of the $\vartheta(\Theta | X)$ in terms of utility maximization. If $Y = T(X)$ is a sufficient statistic for $\theta$, the $\vartheta(\Theta | X) = \vartheta(\Theta | Y)$.

Lindley's work has been influential among Bayesian statisticians concerned with quantification of information (Bernardo 1979a,b; Brooks 1982; DeGroot 1962; Ebrahimi and Soofi 1990; El-Sayyad 1969; Goel 1983; Goel and DeGroot 1979; Soofi 1988, 1990; Stone 1959; Turrero 1989). Some authors have proposed maximization of $\vartheta(\Theta | X)$ for developing priors that add little to the sample information. But because the general solutions are untractable, $\vartheta(\Theta | X)$ has not been used as the main vehicle for developing priors. Lindley (1961) explicated Jeffreys's prior in terms of $\vartheta(\Theta | X)$. He showed that ignorance between two neighboring values $\theta$ and $\theta + \Delta\theta$ implies $\vartheta(\Theta | X) \approx 2(d\theta)^2\mathcal{F}(\theta)$, $\mathcal{F}(\theta)$ being the Fisher information. Bernardo (1979b) developed limiting priors that maximize $\vartheta(\Theta | X)$. Hill and Spall (1987) and Spall and Hill (1990) developed approximate solutions for maximization of $\vartheta(\Theta | X)$ in specific problems.

## 2.4 Jaynes

Jaynes (1957) introduced the *maximum entropy* (ME) *principle* of scientific inference as a generalization of Laplace's "principle of insufficient reason." ME refers to maximizing either $H(\pi)$ or the negative entropy of $\pi$ relative to a dominating measure $p$, $-D(\pi : p)$, with respect to $\pi$. Thus the ME principle extends to Kullback's MDI.

In the ME approach, the partial knowledge about the probability distribution of $X$ is formulated in terms of a set of *information constraints,*

$$E[T_k(X)] = \theta_k, \qquad k = 0, 1, \ldots, K, \qquad (5)$$

where $T_k(X)$ are measurable functions with respect to $d\Pi$ and $\boldsymbol{\theta} = (1, \theta_1, \ldots, \theta_K)'$. The *normalizing constraint* is represented by $T_0(X) = 1$ and $\theta_0 = 1$.

The minimum information model $\pi^*(x | \boldsymbol{\theta})$ with reference to a dominating measure $p(x)$, when it exists, is in the form of

$$\pi^*(x | \boldsymbol{\theta}) = M(\boldsymbol{\beta})p(x)\exp[\beta_1 T_1(x) + \cdots + \beta_K T_K(x)], \qquad (6)$$

where $\beta = (\beta_1, \ldots, \beta_K)'$ is a vector of Lagrange multipliers, $M(\beta)$ is the normalizing factor, and $\theta_k = \partial \log M(\beta)/\partial\beta_k$ (Kullback 1959). The ME model is obtained by letting $p(x) = 1$ in (6).

Shore and Johnson (1980) developed a set of axioms of inductive inference and showed that the ME and MDI methods are uniquely correct for inductive inference when information of type (5) is available. Kullback (1959), Habermann (1984), Rissanen (1986 and 1987), and Csiszar (1991) explored MDI at more abstract levels. Csiszar (1975) and Loh (1985) gave geometric interpretations of MDI. Kapur (1989) discussed extensive applications of the ME method in various areas of engineering and sciences.

The *entropy concentration theorem* (ECT) provides the basic rationale for ME inference. Let $\Omega_\theta$ denote the class of all possible frequency distributions that could be observed in $n$ trials from a distribution $\{\pi(x_j), j = 1, \ldots, J > K\}$ that satisfies constraints (5). Then, asymptotically,

$$H^* - \chi^2(J-K-1, \alpha)/2n \le H[\pi(x|\theta)] \le H^* = H[\pi^*(x|\theta)] = \max_{\pi\in\Omega_\theta} H[\pi(x|\theta)],$$

where $X^2(J-K-1, \alpha)$ is the upper $\alpha$ percentile of the chi-squared distribution with $J-K-1$ degrees of freedom. Details were given by Jaynes (1982). The following example extracted from Jaynes (1982) is illustrative.

*Example 2.1.* Consider predicting frequencies in a large number of rolls of a die. Invoking Laplace's principle of insufficient reason gives the uniform frequencies of $n/6$ for all outcomes in $n$ trials. This is of course in complete agreement with the ME estimate of frequencies when the only available information is the number of outcomes; that is, $\pi^*(x) = \frac{1}{6}$, $x = 1, \ldots, 6$. In addition, the ECT estimates that approximately 95% of all possible frequency distributions have entropies $H(\mathbf{1})$ that satisfy

$$1.792 - 5.503/n \le H(\mathbf{1}) \le H^*(\mathbf{1}) = \log 6 = 1.792. \quad (7)$$

Here $H(\mathbf{1})$ denotes the entropy of all frequency distributions that satisfy the normalizing constraint $T_0(x) = 1$, $H^*(\mathbf{1})$ denotes the entropy of $\pi^*$, and $5.503 = \chi^2(5, .05)/2$. For a modest $n = 100$, (7) gives [1.736, 1.792], which is quite narrow. Thus the ME (uniform) distribution approximates many frequency distributions that may result in the 100 trials.

The ME procedure extends Laplace's principle of insufficient reason for assigning probabilities. For example, if information is given that the die is loaded such that the average outcome is $\theta_1$, then the ME method uses $T_1(x) = x$ and gives the ME probabilities by

$$\pi^*(x_j) = \frac{\exp(\beta x_j)}{\sum_{h=1}^{6} \exp(\beta x_h)}. \quad (8)$$

For example, when $\theta_1 = 4.5$, the Lagrange multiplier is $\beta = .371$ and the ME probabilities are (.054, .079, .114, .165, .240, .348). For $n = 100$, the ECT estimates that approximately 95% of the frequency distributions with mean 4.5 have entropies $H(\mathbf{1}, \mathbf{T}_1)$ such that

$$1.567 \le H(\mathbf{1}, \mathbf{T}_1) \le H^*(\mathbf{1}, \mathbf{T}_1) = 1.614.$$

Again, the ME distribution approximates the outcomes remarkably well.

Extension of the ME procedure to more than one random variable is straightforward. When the constraints used in the ME computation pertain only to the moments of the marginal distributions, then the maximum entropy solution gives independence structure for the stochastic relationship between the marginals (Jaynes 1968). Information about dependency between the variables may be captured by inclusion of product-moment constraints in the ME computations. Gokhale and Kullback (1978) implicitly used this property for testing various dependence structures in contingency tables.

Various ME procedures have appeared in statistics literature in the contexts of developing prior distributions for Bayesian analysis (Berger 1985 and Zellner 1971), goodness-of-fit tests (Arizono and Ohta 1989; Chandra, De Wet, and Singpurwalla 1982; Dudewicz and Van der Meulen 1981; Ebrahimi, Habibullah, and Soofi 1992; Gokhale 1983; Vasicek 1976), and inversion problem (Donoho, Johnstone, Hoch, and Stern 1992 and references therein). But thus far mainly in the Bayesian context, the ME formulation has acquired Jaynes' interpretation of information quantification. Zellner (1971, 1982, 1984) extended this frontier by developing the maximum data information prior, which blends the ME method with Lindley's approach. Zellner's work is currently viewed as the focal point of information-theoretic approach in Bayesian statistics (see Soofi 1995).

## 3. ME INFORMATION DIAGNOSTICS

The ME method is quite versatile for developing models that capture and diagnostics that measure information in various statistical problems (see, for example, Golan, Judge, and Perloff 1993, Mazzuchi, Soofi, and Soyer 1993; Ryu 1993; Soofi, Ebrahimi, and Habibullah 1995; Soofi 1992; Zellner and Min 1992; and references therein). In the ME formulation, the information about prediction of outcomes are due to constraints. We have already seen that in Example 2.1, when only the normalizing constraint **1** was used, the ME probabilities were uniform with entropy $H^*(\mathbf{1}) = 1.792$. Inclusion of the additional constraint $T_1(x) = x$, $E(X) = 4.5$, pulled the ME distribution away from uniformity and pushed the probabilities toward more concentration on the larger outcomes. The inclusion of the additional constraint reduced the maximum entropy to $H^*(\mathbf{1}, \mathbf{T}_1) = 1.614$. The 10% reduction of uncertainty due to the knowledge of $T_1(X)$ quantifies the information content of the additional constraint. This simple idea of quantifying uncertainty reduction provides useful diagnostics for evaluating the merits of explanatory variables in the class of probability models known as *logit*. The die probability model (8) is a simple example of a logit model.

### 3.1 A General Class of Logit Models

Consider the problem of modeling $\{\pi_i = (\pi_{i1}, \ldots, \pi_{iJ})'$ $i = 1, \ldots, n\}$, the probability vectors over a set of events $\{x_1, \ldots, x_J\}$, in terms of explanatory variables $u_{ia} = \mathbf{T}_{ia}(x_j)$

and $v_{ijb} = \mathbf{T}_{ib}(x_j)$, $a = 1, \ldots, A$, $b = 1, \ldots, B$, $j = 1, \ldots, J$. In the terminology of choice modeling, $\pi_{ij}$ is the probability that the $i$th individual selects the $j$th alternative $x_j$ from the choice set $\{x_1, \ldots, x_J\}$, $u_{ia}$ is an attribute of the $i$th individual (e.g., age, income, education), and $v_{ijb}$ is an attribute of the $j$th alternative evaluated by the $i$th individual (e.g., cost, quality).

In the traditional approach, a link function $F$ is assumed that relates the explanatory variables to the probabilities. The form of $F$ is known except for some parameters that are usually estimated by the maximum likelihood method based on the indicator function of the $j$th choice made by the $i$th sampled individual, $y_{ij}$. A widely used link function is logit. In contrast, the ME approach does not require the assumption of a link function, uses the explanatory variables in the formulation of the constraints, and produces the logit model as the solution. Furthermore, the ME model also satisfies some moment information available about the attributes. Let $\mu_{a,j}$ and $\nu_b$ denote the moment parameters associated with the individual-specific attribute $u_{ia}$ and the choice-specific attribute $v_{ijb}$. The parameters $\mu_{a,j}$ and $\nu_b$ may be determined *internally* from the data $y_{ij}$ and/or *externally* based on nonsample information.

The constraints will be formulated as

$$\mathbf{C}'\boldsymbol{\Pi} = \boldsymbol{\theta}. \tag{9}$$

The constituents of (9) are

$$\boldsymbol{\theta}' = [(1, \ldots, 1), \ldots, (1, \ldots, 1), \boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_A', \boldsymbol{\nu}'],$$

where $(1, \ldots, 1)$'s are $n$ vectors of dimension $J$, $\boldsymbol{\mu}_a' = (\mu_{a,1}, \ldots, \mu_{a,J-1})$, and $\boldsymbol{\nu}' = (\nu_1, \ldots, \nu_B)$;

$$\boldsymbol{\Pi}' = [(\pi_{11}, \ldots, \pi_{1J}), \ldots, (\pi_{n1}, \ldots, \pi_{nJ})];$$

and the constraints coefficient matrix $\mathbf{C}$ is constructed as in (10) shown below. The symbols in (10) are defined as follows. The submatrix $\mathbf{1}$ enforces the $n$ normalizing constraints on the $n$ probability vectors. The submatrix $\mathbf{U}$ refers to the set of submatrices $\mathbf{U}_1, \ldots, \mathbf{U}_A$ constructed for the individual-specific attributes $u_{ia}$, $a = 1, \ldots, A$. Because each individual-specific attribute varies only over $i$, inclusion of each attribute requires $J - 1$ constraints, $\mathbf{U}_{aj}'\boldsymbol{\Pi} = \mu_{a,j}$, $j = 1, \ldots, J - 1$. (Inclusion of $J$ constraints makes the rows of $\mathbf{U}_a'$ linearly dependent on rows of $\mathbf{1}$). The submatrix $\mathbf{V}$ is for the alternative-specific attributes, $v_{ijb}$. The alternative-specific attributes vary over all $i$ and $j$; therefore, each attribute requires a single constraint $\mathbf{V}_b'\boldsymbol{\Pi} = \nu_b$, $b = 1, \ldots, B$.

Maximizing the joint entropy of $n$ independent distributions in $\boldsymbol{\Pi}$,

$$\mathbf{H}(\boldsymbol{\Pi}) = -\sum_{i=1}^{n}\sum_{j=1}^{J} \pi_{ij} \log \pi_{ij},$$

subject to (9), gives the ME probabilities

$$\mathbf{C}' = \left[\begin{array}{c} \mathbf{1}' \\ \hline \mathbf{U}' \\ \hline \mathbf{V}' \end{array}\right] = \left[\begin{array}{c} \mathbf{1}' \\ \hline \mathbf{U}_1' \\ \hline \cdot \\ \cdot \\ \cdot \\ \hline \mathbf{U}_A' \\ \hline \mathbf{V}' \end{array}\right] = \left[\begin{array}{c} \mathbf{1}' \\ \hline \mathbf{U}_{1,1}' \\ \cdot \\ \cdot \\ \mathbf{U}_{1,J-1}' \\ \hline \cdot \\ \cdot \\ \cdot \\ \hline \mathbf{U}_{A,1}' \\ \cdot \\ \cdot \\ \mathbf{U}_{A,J-1}' \\ \hline \mathbf{V}_1' \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{V}_B' \end{array}\right] = \left[\begin{array}{ccccccccccccc}
1 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 \\
0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & 1 \\
\hline
u_{11} & 0 & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & u_{n1} & 0 & \cdot & \cdot & 0 \\
0 & u_{11} & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & u_{n1} & \cdot & \cdot & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & \cdot & \cdot & u_{11} & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & u_{n1} & 0 \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
u_{1A} & 0 & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & u_{nA} & 0 & \cdot & \cdot & 0 \\
0 & u_{1A} & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & u_{nA} & \cdot & \cdot & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & \cdot & \cdot & u_{1A} & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & u_{nA} & 0 \\
\hline
v_{111} & \cdot & \cdot & \cdot & v_{1J1} & \cdot & \cdot & \cdot & v_{n11} & \cdot & \cdot & \cdot & v_{nJ1} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
v_{11B} & \cdot & \cdot & \cdot & v_{1JB} & \cdot & \cdot & \cdot & v_{n1B} & \cdot & \cdot & \cdot & v_{nJB}
\end{array}\right]. \tag{10}$$

$$\pi^*_{ij}(\mathbf{U}, \mathbf{V}, \theta) = \frac{\exp\{\alpha'_j(\theta)\mathbf{u}_i + \beta'(\theta)\nu_{ij}\}}{\sum_{h=1}^{J} \exp\{\alpha'_h(\theta)\mathbf{u}_i + \beta'(\theta)\nu_{ih}\}} \quad (11)$$

Here $\mathbf{u}_i = (u_{i1}, \ldots, u_{iA})'$, $\nu_{ij} = (\nu_{ij1}, \ldots, u_{ijB})'$, $\alpha_j(\theta) = (\alpha_{j1}(\theta), \ldots, \alpha_{jA}(\theta))'$, $\beta(\theta) = (\beta_1(\theta), \ldots, \beta_B(\theta))'$, $\alpha_{ja}(\theta)$, $a = 1, \ldots, A$, $j = 1, \ldots, J - 1$, and $\beta_b(\theta)$, $b = 1, \ldots, B$ are Lagrange multipliers, and $\alpha_{aJ} = -\alpha_{a1} - \cdots - \alpha_{a,j-1}$. The ME probabilities (11) are the logit, and the Lagrange multipliers are the logit coefficients.

The ME approach is capable of producing a variety of logit models. Using various forms of constraints and moment parameters $\mu_{a,j}$, and $\nu_b$ in (9) leads to various types of logit models in (11). For example, when only the individual-specific constraints $\mathbf{U}_1, \ldots, \mathbf{U}_A$ are used and the moment parameters are set as $\mu_{aj} = \mathbf{U}'_{aj}\mathbf{y}$ with $\mathbf{y} = [(y_{11}, \ldots, y_{1J})', \ldots, (y_{n1}, \ldots, y_{nJ})']$, the ME solution (11) reduces to the usual logistic regression model, with the Lagrange multipliers being the same as the usual maximum likelihood estimates $\hat{\alpha}_{ja} = \alpha_{ja}(\mathbf{U}'_{aj}\mathbf{y})$. When only the choice-specific attribute constraints $\mathbf{V}$ are used and the moment parameters are set as $\nu_b = \mathbf{V}'_b\mathbf{y}$, the ME solution (11) reduces to the "conditional logit," with the Lagrange multipliers being the same as the usual maximum likelihood estimates of coefficients for the prespecified logit (Soofi 1992). Thus the traditional logit models estimated by the maximum likelihood method constrain the probabilities to satisfy particular types of sample information (e.g., $\mu_{aj} = \mathbf{U}'_{aj}\mathbf{y}$, the number of sampled individuals that belong to a category, say having a college degree, or $\nu_b = \mathbf{V}'_b\mathbf{y}$, the total cost of automobiles purchased by the sampled individuals). But the ME approach is not limited to the use of sample information. The ME formulation is capable of producing logit models that satisfy some nonsample quantities of interest, such as the population proportion of individuals that belong to a category or the average cost of an automobile in the population. Moreover, the corresponding MDI formulations further generalize the logit class (11) to include models that accommodate prior probabilities, $p_{ij}$, directly as

$$\pi^*_{ij}(\mathbf{U}, \mathbf{V}, \theta, p_{ij}) = \frac{p_{ij}\exp\{\alpha'_j(\theta)\mathbf{u}_i + \beta'(\theta)\nu_{ij}\}}{\sum_{h=1}^{J} p_{ih}\exp\{\alpha'_h(\theta)\mathbf{u}_i + \beta'(\theta)\nu_{ih}\}}.$$

The MDI logit is capable of updating, for example, models that are previously developed for the probabilities of the set of alternatives under consideration. In the remainder of this section, I focus on the ME logit (11). All developments can be extended to the MDI logit along the lines of Soofi (1992).

The ME approach, in addition to enabling capture of various forms of sample and nonsample information in the modeling, also provides methods for quantifying the *information value* of the set of information constraints used for developing the model. The information value of the constraints $[(\mathbf{U}, \mu), (\mathbf{V}, \nu)]$ is measured by

$$I^*[(\mathbf{U}, \mu), (\mathbf{V}, \nu)] = \frac{H^*[\mathbf{1}] - H^*[\mathbf{1}, (\mathbf{U}, \mu), (\mathbf{V}, \nu)]}{H^*[\mathbf{1}]}. \quad (12)$$

The information value for each subset of constraints $(\mathbf{U}, \mu)$, $(\mathbf{U}_a, \mu_a)$, $(\mathbf{V}, \nu)$ is measured similarly. For the special case when $\nu = \mathbf{V}'\mathbf{y}$, $I^*[(\mathbf{V}, \nu)]$ becomes the information index defined by Soofi (1992) for the conditional logit model. Thus (12) generalizes previously defined information indices in two directions: inclusion of the individual-specific attributes $\mathbf{U}$ and/or use of arbitrary moment parameters $\mu$ and $\nu$.

In general, the information index $I^*[(\mathbf{U}, \mu), (\mathbf{V}, \nu)]$ is not separable into the information values of the subsets. It is possible, however, to decompose $I^*[(\mathbf{U}, \mu), (\mathbf{V}, \nu)]$ as

$$\begin{aligned} I^*[(\mathbf{U}, \mu), (\mathbf{V}, \nu)] &= I^*[(\mathbf{U}, \mu)] + I^*[(\mathbf{V}, \nu)|(\mathbf{U}, \mu)] \\ &= I^*[(\mathbf{V}, \nu)] + I^*[(\mathbf{U}, \mu)|(\mathbf{V}, \nu)], \end{aligned} \quad (13)$$

where $I^*[(\mathbf{V}, \nu)|(\mathbf{U}, \mu)]$ is the *partial information value* of $(\mathbf{V}, \nu)$, given $(\mathbf{U}, \mu)$, and $I^*[(\mathbf{U}, \mu)|(\mathbf{V}, \nu)]$ is defined similarly. The decompositions in (13) depend on the orders that the constraints are arranged in $\mathbf{C}$. This makes the assessment of the relative importance of each set of attributes problematic (Kruskal 1987; Soofi 1992).

We may wish to further decompose the information value of a subset of constraints into the information values of smaller subsets. Decomposition of $I^*[(\mathbf{U}, \mu)]$ into $I^*[(\mathbf{U}_1, \mu_1)], \ldots, I^*[(\mathbf{U}_A, \mu_A)]$, or decomposition of $I^*[(\mathbf{V}, \nu)]$ into $I^*[(\mathbf{V}_1, \nu_1)], \ldots, I^*[(\mathbf{V}_B, \nu_B)]$, is generally not possible. The partial information method of (13) is possible, but one may also proceed with a decomposition scheme based on orthogonal transformation of the constraints that does not depend on their order of arrangement.

## 3.2 Orthogonal Decomposition of Information

Consider the following approximation of information index $I^*(\mathbf{V}, \nu)$:

$$\begin{aligned} I^*[(\mathbf{V}\ \nu)] &\approx g_{n,J}\beta'\boldsymbol{\Sigma}_{\mathbf{v}}\beta \\ &= g_{n,J}\Big\{\beta_1^2\sigma_{11} + \cdots + \beta_B^2\sigma_{BB} \\ &\qquad + 2\sum_{h<b}\sum \beta_b\beta_h\sigma_{bh}\Big\}, \end{aligned} \quad (14)$$

where $\boldsymbol{\Sigma}_{\mathbf{v}} = [\sigma_{bh}]$ is the variance–covariance matrix of the constraints $\mathbf{V}_1, \ldots, \mathbf{V}_B$, $\beta = \beta(\nu)$ is the vector of Lagrange multipliers for the information constraints, and $g_{n,J}$ depends only on $n$ and $J$. This approximation is based on the equivalence relation $H^*[\mathbf{1}] - H^*[\mathbf{1}, (\mathbf{V}, \nu)] = D[\mathbf{\Pi}^*(\mathbf{V}, \nu):\mathbf{\Pi}^*(\mathbf{1})]$, where $\mathbf{\Pi}^*(\mathbf{1})$ is the set of $n$ uniform distributions (Soofi 1992), and a result given in Gokhale and Kullback (1978, p. 354, eq. (A.15) applied to eq. (37), p. 199).

The purpose of the quadratic approximation (14) is to directly relate predictability of a logit model to the logit coefficients and the variance–covariance structure of the explanatory variables. The relationship (14) indicates that when the explanatory variables are correlated, information is not additively decomposable in terms of the logit coefficients. Therefore, when the explanatory variables are nonorthogonal, the practice of evaluating the saliency of variables by just comparing the magnitudes of the logit coefficients could be misleading.

Any set of $B$ vectors spanning the rows of $\mathbf{V}$'s is information equivalent to $\mathbf{V}_1, \ldots, \mathbf{V}_B$. That is, for any nonsingular transformation $\boldsymbol{\Gamma}$, $I^*[(\mathbf{V}, \boldsymbol{\nu})] = I^*[(\mathbf{V}\boldsymbol{\Gamma}, \boldsymbol{\Gamma}'\boldsymbol{\nu})]$. Let $\boldsymbol{\Gamma}$ be the (orthogonal) matrix of the eigenvectors of $\boldsymbol{\Sigma}_{\mathrm{v}}$ and let $\mathbf{V}\boldsymbol{\Gamma} = \mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_B]$. The transformed constraint set $\mathbf{W}$ is referred to as the *principal components* of $\mathbf{V}$, and $\boldsymbol{\Sigma}_{\mathrm{w}} = \boldsymbol{\Gamma}'\boldsymbol{\Sigma}_{\mathrm{V}}\boldsymbol{\Gamma} = \boldsymbol{\Lambda} = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_K]$, where $\lambda_1, \ldots, \lambda_B$ are the eigenvalues of $\boldsymbol{\Sigma}_{\mathrm{v}}$. Now we can decompose information index $I^*(\mathbf{V}, \boldsymbol{\nu})$ as

$$\begin{aligned} I^*[(\mathbf{V}, \boldsymbol{\nu})] & \\ = I^*[(\mathbf{W}, \boldsymbol{\omega})] &\approx g_{n,J}\boldsymbol{\eta}'\boldsymbol{\Lambda}\boldsymbol{\eta} \\ &= g_{n,J}[\eta_1^2\lambda_1 + \cdots + \eta_B^2\lambda_B] \\ &\approx I^*[(\mathbf{W}_1, \omega_1)] + \cdots + I^*[(\mathbf{W}_B, \omega_B)], \end{aligned} \quad (15)$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_B)' = \boldsymbol{\Gamma}'\boldsymbol{\nu}$ and $\boldsymbol{\eta} = \boldsymbol{\Gamma}'\boldsymbol{\beta}$. Note that $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_B)'$ is the vector of Lagrange multipliers (i.e., logit coefficients) for the orthogonalized constraints (i.e., variables).

In (15), $I^*[(\mathbf{W}_k, \omega_k)] \approx g_{n,J}\eta_k^2\lambda_k$. Thus the information value of an orthogonalized constraint is proportional to the product of the constraint variance $\mathrm{var}(\mathbf{W}_k) = \lambda_k$ and the square of its logit coefficient $\eta_k^2$. This is of course quite in accord with our intuition that an explanatory variable with a high variance should be more informative than one with a low variance and that a large logit coefficient indicates a high "impact" on the response. The information quantity $I^*[(\mathbf{W}_k, \omega_k)]$ combines both of these intuitive elements into a single criterion.

The relative importance of an orthogonalized component $\mathbf{W}_k$ is assessed by the proportion of total information contained in $(\mathbf{W}_k, \omega_k)$,

$$\varphi[(\mathbf{W}_k, \omega_k)] = \frac{I^*[(\mathbf{W}_k, \omega_k)]}{I^*[(\mathbf{V}, \boldsymbol{\nu})]} \approx \frac{\eta_k^2\lambda_k}{\eta_1^2\lambda_1 + \cdots + \eta_B^2\lambda_B}, \qquad k = 1, \ldots, B. \quad (16)$$

Within each set of the individual-specific constraints $\mathbf{U}_a$, the information is further decomposable into the information values of the single constraints, $I^*[(\mathbf{U}_a, \boldsymbol{\mu}_a)] = I^*[(\mathbf{U}_{a,1}, \mu_{a,1})] + \cdots + I^*[(\mathbf{U}_{a,J-1}, \mu_{a,J-1})]$, because for each $a = 1, \ldots, A$, $\mathbf{U}_{a,1}, \ldots, \mathbf{U}_{a,J-1}$ or orthogonal. Furthermore, because $\mathrm{var}(\mathbf{U}_{a,1}) = \cdots = \mathrm{var}(\mathbf{U}_{a,J-1})$, we can measure the relative importance of the $a$th individual-specific attribute for the $j$th choice by

$$\varphi[(\mathbf{U}_{a,j}, \mu_{a,j})] = \frac{I^*[(\mathbf{U}_{a,j}, \mu_{a,j})]}{I^*[(\mathbf{U}_a, \boldsymbol{\mu}_a)]} \approx \frac{\alpha_{a,j}^2}{\alpha_{a,1}^2 + \cdots + \alpha_{a,J-1}^2}, \qquad j = 1, \ldots, J-1.$$

Transformation of a set of explanatory variables in the direction of principal components is very popular in applied fields. It is quite common first to reduce the number of variables to a few principal components selected based on the eigenvalues of the covariance matrix of the explanatory variables and then to use the chosen components in a subsequent model. A number of authors have noted that in regression analysis, such a two-step procedure may result in exclusion of components that could have high explanatory power in the subsequent model. To avoid this problem, principal component regression methods have been proposed that combine the two criteria for selecting a number of components and estimating a regression model together (see Soofi 1988 for an information-theoretic approach). From (16) we see that principal components selection based on the eigenvalues $\lambda_k$ for use in a subsequent logit analysis may result in nonnegligible loss of information when the coefficient $\eta_k$ of a component $\mathbf{W}_k$ is relatively large and the corresponding eigenvalue $\lambda_k$ is relatively small. The information diagnostic (16) incorporates both $\eta_k$ and $\lambda_k$. Interestingly, an information index developed for principal component linear regression by Soofi (1988) based on Lindley's measure also combines eigenvalue $\lambda_k$ with the mean squared error of regression $s_k^2$.

## 3.3 Visualizing ME Information

The following simple example helps to visualize the concept of information in ME modeling.

*Example 3.1.* Suppose that there are three numerical outcomes $x_j$ with probabilities $\pi_j$, $j = 1, 2, 3$, and suppose that $x_3$ is the average outcome, so $\theta_1 = x_3$. That is, $x_1 < x_3 < x_2$ or $x_2 < x_3 < x_1$. The information constraint is defined by $\mathbf{T}_1\boldsymbol{\pi} = x_3$, with $\mathbf{T}_1(x_j) = x_j$. Figure 5 shows a number of possible information constraints on the $(\pi_1, \pi_2)$ plane defined by lines $\pi_2 = \alpha\pi_1$, with $\alpha = (x_1 - x_3)/(x_3 - x_2)$. Also shown in Figure 5 are the foci of the ME probabilities given by

$$\pi_j^* = \frac{\exp(\beta x_j)}{\sum_{h=1}^3 \exp(\beta x_h)}, \qquad \beta = \frac{\log \alpha}{x_2 - x_1}.$$

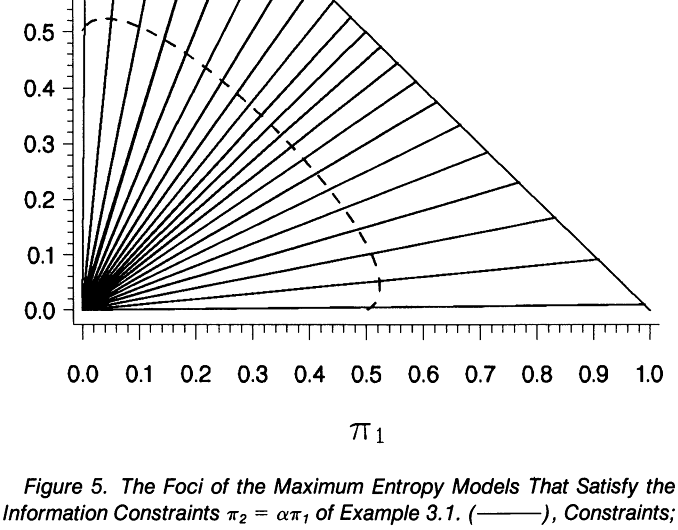

*Figure 5. The Foci of the Maximum Entropy Models That Satisfy the Information Constraints $\pi_2 = \alpha\pi_1$ of Example 3.1. (———), Constraints; (-----), ME Models.*

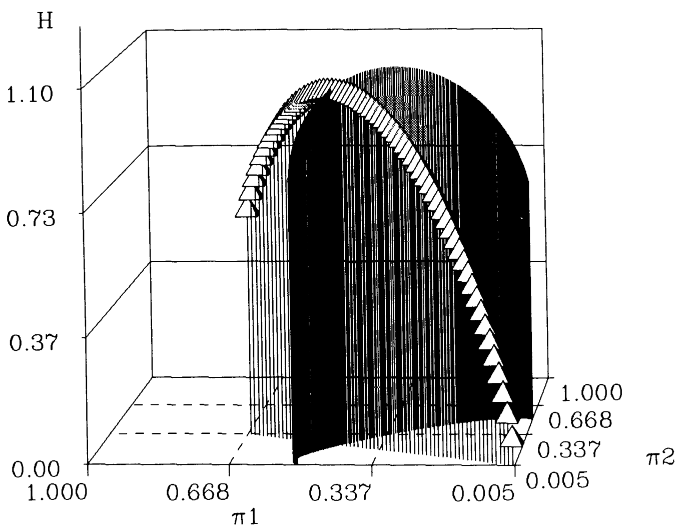

*Figure 6. The Entropies of Maximum Entropy Models that Satisfy Constraints $\pi_2 = \alpha\pi_1$ (Shaded Bars) and the Entropies of Models that Satisfy a Single Constraint $\pi_2 = .5\pi_1$ (Arrow-Headed Bars), Discussed in Example 3.1.*

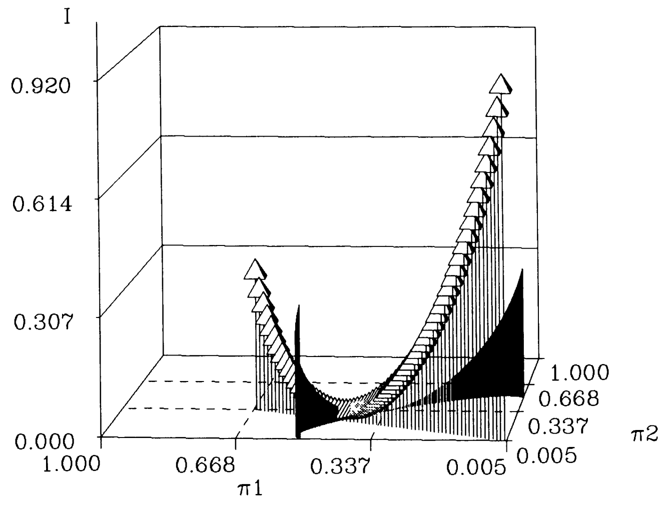

*Figure 7. The Information in Maximum Entropy Models That Satisfy Constraints $\pi_2 = \alpha\pi_1$ (Shaded Bars) and the Information in the Models That Satisfy a Single Constraint $\pi_2 = .5\pi_1$ (Arrow-Headed Bars), Discussed in Example 3.1.*

The least informative constraint is given by $\alpha = 1$ (obtained when $x_3 = (x_1 + x_2)/2$), so $\pi^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $I^*(\mathbf{T}_1) = 0$. The constraints are most informative when $\alpha \approx 0$ (obtained by $x_1 \approx x_3$) or when $\alpha \approx \infty$, (obtained by $x_1 \approx x_2$). The most informative ME probabilities are $\pi^* \approx (\frac{1}{2}, 0, \frac{1}{2})$ or $\pi^* \approx (0, \frac{1}{2}, \frac{1}{2})$. The most informative information index is $I^*(\mathbf{T}_1) \approx 1 - \log(2)/\log(3) = .369$. Figure 6 shows the entropies of the family of distributions satisfying a particular constraint (arrow-headed bars) and the entropies of the ME models in the families satisfying various constraints (shaded bars). The corresponding information quantities are shown in Figure 7.

Note that in this example the most informative constraints can reduce the maximum uncertainty over a three-outcome problem to the maximum uncertainty over a two-outcome problem. This idea is useful for calibrating information indices. In general, we may calibrate an $I^*(\mathbf{T})$ by the information index of a hypothetical model that is able to reduce the number of outcomes from $J$ to $J'$ given by

$$I_{J,J'} = 1 - \frac{\log(J')}{\log(J)}, \qquad J' = 1, \ldots, J - 1.$$

We find $J'$ such that $I^*(\mathbf{T}) \approx I_{J,J'}$. We then conclude that $\mathbf{T}$ is able to reduce our uncertainty about occurrence of $J$ outcomes to the uncertainty about occurrence of $J'$ outcomes, $J' \le J$. In the die example of Section 2.4, the information value of the constraint $\mathbf{T}_1(x) = x$ with $\theta_1 = 4.5$ is $I^*(\mathbf{T}_1) = .1$. Noting that $I_{6,5} = 1 - \log(5)/\log(6) = .1$, the interpretation of $I^*(\mathbf{T}_1) = .1$ is that the constant is able to reduce a six-outcome problem to an almost five-outcome problem.

## 4. INFORMATION AND CHAOS

In statistical information theory, the notion of randomness plays the central role. The notion of randomness is of course not so well-defined and commonly is contrasted with deterministic behavior. But some purely deterministic dynamical systems, referred to as *chaos,* show nondeterministic features in any practical sense. Wegman (1988) demonstrated that outcomes of chaotic systems that from a mathematical viewpoint are purely deterministic also satisfy conditions put forward by Cramer (1946) and de Finetti (1974) for defining randomness. In chaos literature, deterministic chaos and randomness are intertwined by the ergodic theory. Also evident in chaos literature are endeavors to quantify chaotic uncertainty by a number of information functions. Entropy has been used in chaos literature (Ruelle 1989); however, discrimination information and mutual information have not yet been utilized in the context of chaos. Recently, Berliner (1991) considered the use of Fisher's information in analysis of noisy chaotic data. This section presents a discrimination information diagnostic and shows that in the context of chaos, the discrimination information interpretation of Fisher's information is more meaningful than its usual statistical interpretation.

Mathematics of chaos centers on deterministic dynamic systems such as difference equations,

$$x_{t+1} = F(x_t), \qquad t = 0, 1, 2, \ldots,$$

where $F$ is a nonlinear function. The outcome at any time $t$ depends on the initial condition $x_0$ by recursion $x_t = F(F(\ldots(F(x_0)))) = F^t(x_0)$. Sensitivity of an outcome to initial condition is considered the most important characteristic of chaotic behavior. Under appropriate conditions, the system's sensitivity to the initial condition is summarized by the characteristic (Liapunov) exponent,

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \log\left|\frac{dF^t(x_0)}{dx_0}\right|.$$

For a chaotic system, $\lambda$ is positive.

Under a chaotic regime with an approximately exponential rate of growth, $dx_t \approx dx_0\exp(\lambda t)$, seemingly indistinguishable initial conditions evolve into distinguishable points. In chaos literature such an impact has been interpreted as information created by the chaotic map $F$ about the initial conditions. Entropy has been suggested for quantifying the rate of information creation. Suppose that $F$ is measure-preserving, $P[F(x)] = P(x)$, with respect to an ergodic distribution $P$ over a compact phase space $X$. An estimate of $H(P)$ is found as follows. The phase space $X$ is partitioned into $J(\varepsilon)$ pieces of size $\varepsilon$ each with probability $P_j(\varepsilon)$, $j = 1, \ldots, J$. (Here $\varepsilon$ signifies a given measurement precision, so that points within a partition are not distinguishable.) Then $P_j(\varepsilon)$ is estimated by its multinomial approximation $\pi_j(\varepsilon)$. Finally, the information content of $P$ for a given partition $X$ is estimated by the entropy $H_t[\pi(\varepsilon)]$, where $\pi(\varepsilon)$ is the vector of the multinomial probabilities (see, for example, Berliner 1992). The *information dimension* is defined by

$$\xi(P) = -\lim_{\varepsilon \to 0} \frac{H[\pi(\varepsilon)]}{\log(\varepsilon)}. \qquad (17)$$

The numerator in (17) is referred to as Kolmogorov–Sinai entropy. There are some other quantities referred to as entropy in the chaos literature (see Ruelle 1989), but the information dimension is the most well known of all.

The issue of sensitivity to initial condition naturally lends itself to the problem of discrimination between $x_0$ and $x_0 + \Delta x_0$ based on a set of data. Consider a dynamic model with noise,

$$Y_t = x_t + e_t, \qquad x_t = F(x_{t-1}; \xi), \qquad e_t\text{'s IID } \pi(e_t; \theta), \qquad (18)$$

where $\xi$ is a vector of parameters of the dynamic system and $\theta$ is a vector of parameters for the distribution of the noise.

Let $\pi_1(y_t) = \pi[y_t; F^t(x_0 + \Delta x_0, \xi), \theta]$ and $\pi_2(y_t) = \pi[y_t; F^t(x_0, \xi), \theta]$. Then $D(\pi_1 : \pi_2)$ quantifies the mean amount of information contained in $y_t$ for discrimination between $\pi_1$ and $\pi_2$, which differ only in their dependency on $x_0$ and $x_0 + \Delta x_0$. The distinguishability of noisy chaotic systems like (18) may be measured by the discrimination information exponent

$$\zeta = \lim_{\Delta x_0 \to 0} \frac{[D(\pi_1 : \pi_2)]^{1/2}}{\Delta x_0}. \qquad (19)$$

Berliner (1991) considered the model (18) with Gaussian noise $\pi(e_t; \sigma) = N(0, \sigma^2)$. The Gaussian likelihood function based on $y_1, \ldots, y_n$ is

$$L(x_0, \xi, \sigma) = \frac{1}{[(2\pi)^{1/2}\sigma]^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=1}^{n} [y_t - F^t(x_0; \xi)]^2\right\}.$$

The Fisher information for $x_0$ based on the Gaussian likelihood is

$$\mathcal{F}(x_0) = \frac{1}{\sigma^2} \sum_{t=1}^{n} \left(\frac{dx_t}{dx_0}\right)^2.$$

Berliner observed that $F(x_0)$ is approximately proportional to the exponential rate of growth for $F$, $\exp[2n\lambda(\xi)]$, when $n$ is large and $\lambda(\xi) > 0$. From this Berliner (1991) concluded that "clearly both Liapunov exponent (when it is positive) and Fisher information measure similar characteristics of the model as $n$ tends to infinity. However, the value we attach to large Fisher information is different for chaotic processes than the more familiar interpretation for 'nicer' problems, such as problems with the monotone likelihood ratio property" (p. 942). The familiar interpretation for "nicer" problems refers to the precision with which a parameter of a statistical model can be estimated—in Lehmann's words, "the ease with which a parameter can be estimated" (Lehmann 1983, p. 120).

For the Gaussian noise case, the expected information in $y_t$ for discrimination between two distributions is

$$D(\pi_1 : \pi_2) = \frac{[F^t(x_0 + \Delta x_0, \xi) - F^t(x_0; \xi)]^2}{2\sigma^2}.$$

Upon division by $(\Delta x_0)^2$ and taking limit, we will observe the agreement between the discrimination information and the Liapunov exponent, both measuring the distinguishability aspect of $F$. This is an example of the interpretation of Fisher's information as an approximation to information contained in data for discrimination between two distributions in the same parametric family that differ infinitesimally over a parameter space. This is in fact exactly the agreement between the Fisher information under Gaussian error and Liapunov exponent observed by Berliner. This suggests that the discrimination information exponent defined in (19) could be viewed as a generalization of the Liapunov exponent.

## 5. FUTURE DIRECTIONS

The current research in information-theoretic statistics is concerned mainly with developing information-theoretic models that capture and diagnostics that quantify information in various statistical problems. In this article I have presented examples from two diverse fields: choice modeling and chaos. The ME approach to modeling and diagnosis provides research opportunities in a number of directions. The ME diagnostics developed for the logit analysis provide the basis for developing various modeling tools for logit analysis analogous to those available for linear regression analysis. The orthogonal decomposition (15) provides a basis for studying collinearity and related problems in developing logit models. For example, one may now consider developing principal component logit estimation procedures analogous to the principal component regression methods.

The discrimination information diagnostic (19) proposed for studying chaos is quite general. It needs to be more thoroughly examined in analysis of chaotic data under non-Gaussian noise models. The question of unpredictability of $x_{t+s}$ based on $x_t$ (for large $s$) despite the exact mathematical relationship $x_{t+s} = F^s(x_t)$ can be studied in terms of the mutual information. Developing mutual information diagnostics requires extending conditions under which Kolmogorov–Sinai entropy is defined to the two-dimensional phase space $X_t \times X_{t+s}$. The mutual information may also be utilized according to Lindley's approach in the Bayesian analysis of chaos by extending the work of Berliner (1991) and Geweke (1992).

Many papers that have developed information diagnostics for various statistical models have been cited. There are as yet some important problems in statistics and related fields for which developing diagnostics that quantify information would be useful. For example, consider the following problem (posed by Wegman in personal conversation): When an analyst uses a parametric model for statistical analysis, he or she introduces extraneous (i.e., nonsample) information and often obtains more efficient estimates than one obtains using a nonparametric method. It would be interesting to develop diagnostics that enable the analyst to quantify the additional amount of information injected in the analysis by a model. Such information diagnostics will be extremely useful, because they will help analysts to compare the tradeoff between the strength of the assumptions made via the model and the gain of estimation efficiency.

In general, developing an ME model and performing inference based on ME diagnostics require extensive computation. The use of the ME diagnostics beyond the descriptive level and incorporating uncertainty involved in the ME diagnostics require efficient computational algorithms. In the sampling theory approach, the uncertainty associated with an ME diagnostic may be accounted for by using Monte Carlo techniques such as bootstrap. Taking uncertainty into account in the Bayesian approach will also be computationally intensive. The ME formulation provides an opportunity for very realistic Bayesian analyses, because often the moment parameters have "physical" meaning. When the moment parameters ($\mu$ and $\nu$) are set externally, one may begin with developing priors for $\mu$ and $\nu$, then derive the priors for the model parameters $\beta(\mu, \nu)$ and other parametric functions of interest, such as the ME diagnostics. Success of ME modeling in practice will depend largely on developing user-friendly codes that can execute these tasks.

Geometrical representations of information-theoretic procedures are very helpful in developing intuition about the ME quantities. The use of abstract geometry is quite important for researchers in the field, and easily understood geometry for the users of information-theoretic methods is needed as well. We have seen that modeling a three-dimensional probability vector with a single information constraint, in addition to the normalizing constraint, exhausts the three dimensions of the perpendicular coordinate systems. Thus visualizing information in a serious modeling problem requires developing other systems, such as the parallel coordinate systems along the line of Wegman (1990).

*[Received March 1993. Revised November 1993.]*

## REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory,* eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267-281.

Arizono, I., and Ohta, H. (1989), "A Test for Normality Based on Kullback-Leibler Information," *The American Statistician,* 34, 20-23.

Ash, R. B. (1965), *Information Theory,* New York: Dover.

Berliner, L. M. (1992), "Statistics, Probability and Chaos" (with discussion), *Statistical Science,* 7, 69-90.

——— (1991), "Likelihood and Bayesian Prediction of Chaotic Systems," *Journal of the American Statistical Association,* 86, 938-952.

Bernardo, J. M. (1979a), "Expected Information as Expected Utility," *The Annals of Statistics,* 7, 686-690.

——— (1979b), "Reference Posterior Distributions for Bayesian Inference" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 41, 113-147.

Bozdogan, H. (1990), "On the Information-Based Measure of Covariance Complexity and Its Application to the Evaluation of Multivariate Linear Models," *Communications in Statistics, Part A—Theory and Methods,* 19, 221-278.

Brockett, P. L. (1991), "Information-Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications," (with discussion) *Transactions of the Society of Actuaries,* 43, 73-135.

Brooks, R. J. (1982), "On the Loss of Information Through Censoring," *Biometrika,* 69, 137-144.

Chandra, M., De Wet, T., and Singpurwalla, N. D. (1982), "On the Sample Redundancy and a Test for Exponentiality," *Communications in Statistics, Part A—Theory and Methods,* 11, 429-438.

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory,* New York: John Wiley.

Cramer, H. (1946), *Mathematical Method of Statistics,* Princeton, NJ: Princeton University Press.

Csiszar, I. (1991), "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference in Linear Inverse Problems," *The Annals of Statistics,* 19, 2032-66.

——— (1975) "I-Divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability,* 3, 146-158.

de Finetti, B. (1970), *Theory of Probability,* New York: John Wiley.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992), "Maximum Entropy and Nearly Black Object" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 54, 41-81.

Dudewicz, E. J., and Van Der Meulen, E. C. (1981), "Entropy-Based Tests of Uniformity," *Journal of the American Statistical Association,* 76, 967-974.

Ebrahimi, N., and Soofi, E. S. (1990), "Relative Information Loss Under Type II Censored Exponential Data," *Biometrika,* 77, 429-35.

Ebrahimi, N., Habibullah, M., and Soofi, E. S. (1992), "Testing Exponentiality Based on Kullback-Leibler Information," *Journal of the Royal Statistical Society,* Ser. B, 54, 739-748.

El-Sayyad, G. M. (1969), "Information and Sampling From Exponential Distribution," *Technometrics,* 11, 4145.

Fisher, R. A. (1921), "On Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London,* Ser. A, 222, 309-368.

Geweke, J. (1992), "Comment: Inference and Prediction in the Presence of Uncertainty and Determinism," *Statistical Science,* 7, 94-101.

Goel, P. K. (1983), "Information Measures and Bayesian Hierarchical Models," *Journal of the American Statistical Association,* 78, 408-410.

Goel, P. K., and DeGroot, M. H. (1979), "Comparison of Experiments and Information Measures," *The Annals of Statistics,* 7, 1066-1077.

Gokhale, D. V. (1983), "On Entropy-Based Goodness-of-Fit Tests," *Computational Statistics and Data Analysis,* 1, 157-165.

Gokhale, D. V., and Kullback, S. (1978), *The Information in Contingency Tables,* New York: Marcel Dekker.

Golan, A., Judge, G., and Jeffrey, P. (1993), "Recovering Information From Multinomial Response Data," unpublished manuscript, University of California, Berkeley.

Good, I. J. (1950), *Probability and the Weighing of Evidence,* London: Griffin.

Haberman, S. J. (1984), "Adjustment by Minimum Discriminant Information," *The Annals of Statistics,* 12, 971-988.

Hill, S. D., and Spall, J. C. (1987), "Noninformative Bayesian Priors for Large Samples Based on Shannon Information," in *Proceedings of the IEEE Conference on Decision and Control,* New York: Institute of Electrical and Electronics Engineers, 1690-1693.

Jaynes, E. T. (1982), "On the Rationale of Maximum-Entropy Methods," *Proceedings of IEEE,* 70, 939-952.

——— (1968), "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics,* SSC-4, 227-241.

——— (1957), "Information Theory and Statistical Mechanics," *Physical Review,* 106, 620-630.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings Royal Society London,* Ser. A, 186, 453-461.

Johnson, W. (1987), "The Detection of Influential Observations for Allocation, Separation, and the Determination of Probabilities in a Bayesian Framework," *Journal of Business & Economic Statistics,* 5, 369-381.

Johnson, W., and Geisser, S. (1982), "Assessing the Predictive Influence of Observations," in *Essays in Honor of C. R. Rao,* eds. Kallianpur, Krishnaiah, and Ghosh, Amsterdam: North-Holland.

——— (1983), "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis," *Journal of the American Statistical Association,* 78, 137-144.

Kapur, J. N. (1989), *Maximum Entropy Models in Science and Engineering,* New York: John Wiley.

Kruskal, W. (1987), "Relative Importance by Averaging Over Orderings," *The American Statistician,* 41, 6–10.

Kullback, S. (1959), *Information Theory and Statistics,* New York: John Wiley (reprinted in 1968 by Dover).

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics,* 22, 79–86.

Lindley, D. V. (1956), "On a Measure of the Information Provided by an Experiment," *The Annals of Mathematical Statistics,* 27, 986–1005.

——— (1961), "The Use of Prior Probability Distributions in Statistical Inference and Decision," *Proceedings of the Fourth Berkeley Symposium,* 1, 436–468.

Loh, W. (1985), "A Note on the Geometry of Kullback-Leibler Information Numbers," *Communications in Statistics, Part A—Theory and Methods,* 14, 895–904.

Mazzuchi, T. A., Soofi, E. S., and Soyer, R. (1993), "Information Diagnostics for Bayesian Reliability Analysis," *Proceedings of the Section on Bayesian Statistical Science of the American Statistical Association,* pp. 173–178.

McCulloch, R. E. (1989), "Local Model Influence," *Journal of the American Statistical Association,* 84, 473–478.

Rissanen, J. (1987), "Stochastic Complexity" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 49, 223–239.

——— (1986), "Stochastic Complexity and Modeling," *The Annals of Statistics,* 14, 1080–1100.

Ruelle, D. (1989), *Chaotic Evolution and Strange Attractors: The Statistical Analysis of Time Series for Deterministic Nonlinear Systems,* Cambridge, U.K.: Cambridge University Press.

Ryu, H. K. (1993), "Maximum Entropy Estimation of Density and Regression Functions," *Journal of Econometrics,* 56, 397–440.

Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica,* 46, 1273–1292.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal,* 27, 379–423.

Shore, J. E., and Johnson, R. W. (1980), "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions on Information Theory,* IT-26, 26–37.

Soofi, E. S. (1988), "Principal Component Regression Under Exchangeability," *Communications in Statistics, Part A—Theory and Methods,* 17, 1717–1733.

——— (1990), "Effects of Collinearity on Information About Regression Coefficients," *Journal of Econometrics,* 43, 255–274.

——— (1992), "A Generalizable Formulation of Conditional Logit With Diagnostics," *Journal of the American Statistical Association,* 87, 412–816.

——— (1995), "The Information Theory and Bayesian Statistics," in *Bayesian Statistics and Applications: A Tribute to Arnold Zellner,* eds. D. Berry, K. Chaloner, and J. Geweke, New York: John Wiley.

Soofi, E., Ebrahimi, N., and Habibullah, M. (1995), "Information Distinguishability With Application to Analysis of Failure Data," submitted to *Journal of the American Statistical Association.*

Soofi, E. S., and Gokhale, D. V. (1991), "Minimum Discrimination Information Estimator of the Mean With Known Coefficient of Variation," *Computational Statistics and Data Analysis,* 11, 165–177.

Spall, J. C., and Hill, S. D. (1990), "Least-Informative Bayesian Prior Distributions for Finite Samples Based on Information Theory," *IEEE Transactions on Automatic Control,* 35, 580–583.

Stone, M. (1959), "Application of a Measure of Information to the Design and Comparison of Regression Experiments," *The Annals of Mathematical Statistics,* 30, 55–70.

Turrero, A. (1989), "On the Relative Efficiency of Grouped and Censored Survival Data," *Biometrika,* 76, 125–131.

Vasicek, O. (1976), "A Test for Normality Based on Sample Entropy," *Journal of the Royal Statistical Society,* Ser. B, 38, 54–59.

Wegman, E. J. (1988), "On Randomness, Determinism and Computability," *Journal of Statistical Planning and Inference,* 20, 279–294.

——— (1990), "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of the American Statistical Association,* 85, 664–675.

Young, A. S. (1987), "On the Information Criterion for Selecting Regressors," *Metrika,* 34, 185–194.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics,* New York: John Wiley (reprinted in 1987 by Krieger, Malabar, FL).

——— (1982), "On Assessing Prior Distributions and Bayesian Regression Analysis With *g*-Prior Distributions," in *Bayesian Decision Techniques: Essays in Honor of Bruno de Finetti,* eds. P. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233–243.

——— (1984), *Basic Issues in Econometrics,* Chicago: University of Chicago Press.

Zellner, A., and Min, C. (1992), "Bayesian Analysis, Model Selection and Prediction," invited paper presented at a symposium in honor of E. T. Jaynes, H. G. B. Alexander Research Foundation, Graduate School of Business, University of Chicago.