I^2SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2009-9
July 23, 2009

Information Measures in Perspective

Nader Ebrahimi
Division of Statistics
Northern Illinois University

Ehsan S. Soofi
Sheldon B. Lubar School of Business
University of Wisconsin-Milwaukee

Refik Soyer
Department of Decision Sciences
The George Washington University

To appear in the International Statistical Review

Information Measures in Perspective

Nader Ebrahimi
Division of Statistics
Northern Illinois University
DeKalb, IL 60155
nader@math.niu.edu

Ehsan S. Soofi
Sheldon B. Lubar School of Business
University of Wisconsin-Milwaukee
P.O.Box 742, Milwaukee, WI 53201
esoofi@uwm.edu

Refik Soyer
Department of Decision Sciences
George Washington University
Washington D.C. 20052
soyer@gwu.edu

February 12, 2010

Abstract

Information-theoretic methodologies are increasingly being used in various disciplines. Frequently an information measure is adapted for a problem, yet the perspective of information as the unifying notion is overlooked. We set forth this perspective through presenting information-theoretic methodologies for a set of problems in probability and statistics. Our focal measures are Shannon entropy and Kullback-Leibler information. The background topics for these measures include notions of uncertainty and information, their axiomatic foundation, interpretations, properties, and generalizations. Topics with broad methodological applications include discrepancy between distributions, derivation of probability models, dependence between variables, and Bayesian analysis. More specific methodological topics include model selection, limiting distributions, optimal prior distribution and design of experiment, modeling duration variables, order statistics, data disclosure, and relative importance of predictors. Illustrations range from very basic to highly technical ones that draw attention to subtle points.

Key Words: Bayesian information, Dynamic information, Entropy, Kullback-Leibler information, Mutual information.

1 Introduction

The information theory offers measures that have axiomatic foundation and are capable of handling diverse problems in a unified manner. However, the information methodologies are often developed in isolation, where a particular measure is used without consideration of the larger picture. This paper takes an integrative approach and draws the common properties of various information measures. This approach enables us to relate solutions to diverse problems. We present information-theoretic solutions to several statistical problems, ranging from very basic to highly technical.

The paper is organized into the following parts:

(a) Basic notions and measures of uncertainty and information (Sections 2-4).

(b) Information methodologies for model derivation and selection, measuring dependence, and Bayesian analysis (Sections 5-7).

(c) Specific areas of applications (Section 8).

The presentations are in terms of random variables and vectors are used only when needed.

The first part of the paper begins with an overview of the foundation literature on concepts of uncertainty and information in Section 2. From the literature, we decipher that two desirable properties of uncertainty measures are concavity and attaining global maximum at the uniform distribution. We illustrate that the variance, although being concave, is not a satisfactory measure for mapping uncertainty. We also conclude that two desirable properties of information measures are convexity and non-negativity. Our focal uncertainty and information measures are, respectively, Shannon entropy (Shannon, 1948) and KL information (Kullback and Leibler, 1951). Sections 3 and 4 present these two key measures, some of their properties, and a few of their generalizations.

In the second part of the paper, Section 5 presents the maximum entropy (Jaynes, 1957) and minimum discrimination information criteria (Kullback, 1959) for derivation of probability models based on moment constraints. This section includes the geometric interpretation (Csiszar, 1975, 1991) and the axiomatic foundation (Shore and Johnson, 1980) of these criteria. This section also points out the roles of the maximum entropy characterization of probability models in methodological problems such as Akaike information criteria (Akaike, 1973) for model selection and the Central Limit Theorem. Section 6 addresses the problem of information about the stochastic dependence between variables. Definitions of independence in terms of the conditionals, marginals, and joint distributions lead to four formulations of dependence information. In terms of Shannon entropy and KL information, all four formulations give the same measure, known as the mutual information (Shannon, 1948). But this is not true for the generalizations of Shannon entropy and KL information. A simple example illustrates use of the mutual information as a tool to show that lower-order independence between variables does not imply mutually independence. This example also shows that the correlation coefficient cannot detect nonlinear dependence. A multivariate normal example illustrates the mutual information as a function of the correlation parameter. Other examples show the usefulness of the information measure when the correlation is not defined. The measure of sample information about parameter introduced by Lindley (1956) has been influential in the development of Bayesian concepts and methods. Section 7 presents an overview of Lindley's measure, its interpretation as a utility function (Bernardo, 1979a), and its methodological applications, including criteria for developing prior distribution, model for likelihood function, optimal design, and regression diagnostics. This section also includes the criterion for the maximal data information prior proposed by Zellner (1977). We also note that due to its KL representation, the "information processing rule" of Zellner (1988, 2002) is endowed with the same axiomatic foundation.

In the final part, Section 8 presents four areas of applications and gives additional references in statistics and some related fields. The first two areas are duration analysis and order statistics, where use of information measures are evolving rapidly. For the duration analysis, we present dynamic information measures which are functions of the age, and dynamic information criteria for developing probability models subject to hazard rate constraints. For the order statistics we briefly state some properties. The final two interdisciplinary areas of applications are data disclosure and relative importance of predictors. Information measures have been used in these areas, which offer ample opportunities for future research. This part concludes with additional references in statistics, econometrics, engineering, and computational biology.

2 Notions of Uncertainty and Information

The notions of uncertainty and information are relative and involve comparison of distributions. Figure 1 illustrates these two concepts through four pairs of probability density functions (pdfs). We shall specify the parameters of these distributions as we proceed. For the present, we seek answers to the following questions visually: Which of the two distributions in each pair displays a set of outcomes that one can predict with a *higher probability*? By which distribution in each pair is it *less difficult* to predict the outcomes? Visually, the answer is the distributions shown as solid curves. These are more concentrated to great extents on some subsets of outcomes than their counterparts shown as dashed curves. An uncertainty function, $\mathcal{U}(f)$, maps each of these distributions to a real number such that in each panel $\mathcal{U}(f_s) < \mathcal{U}(f_d)$, where s denotes those shown as solid curves and d denotes their counterparts in the Panels. Such an $\mathcal{U}(f)$ enables us to rank the predictability of all eight distributions in Figure 1. An information discrepancy function maps

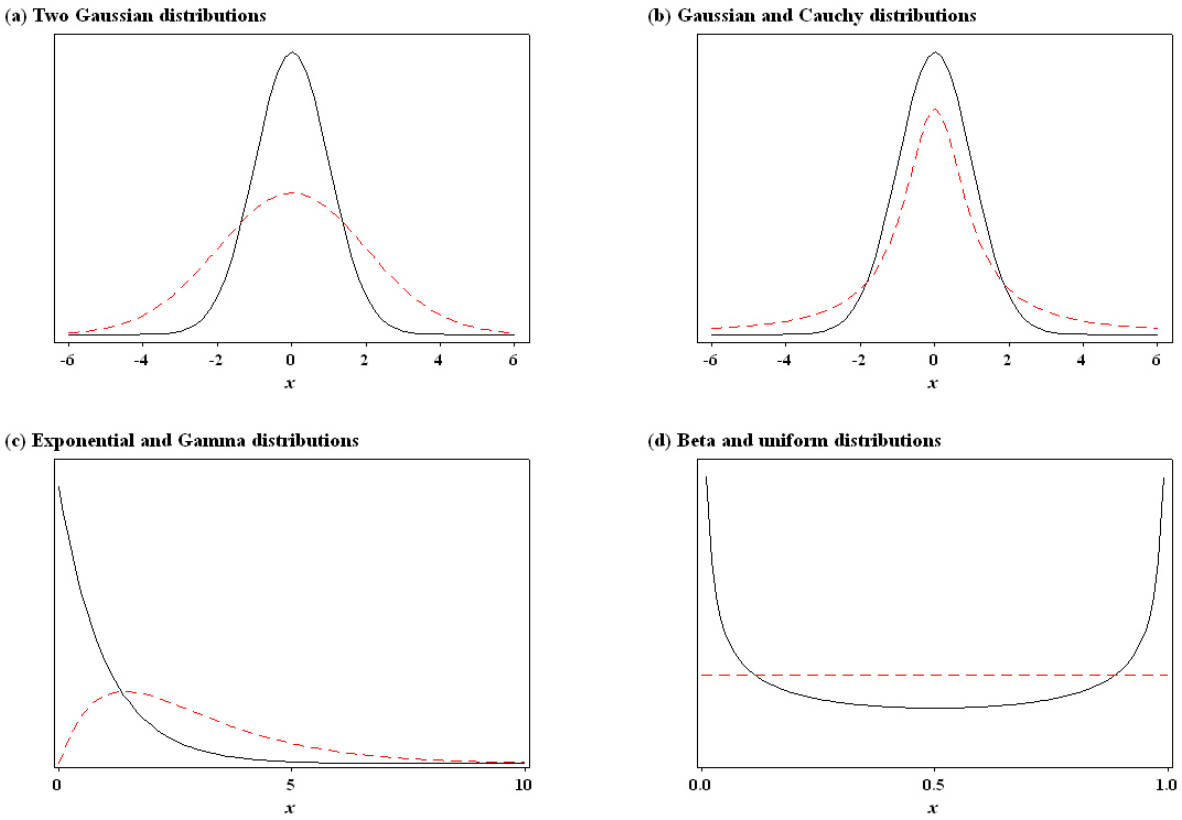

Figure 1: Probability density functions with various levels of concentration.

each pair of these distributions to a nonnegative number, which enables us to evaluate the gain or loss of using one of the distributions in the pair instead of the other.

2.1 Notion of Uncertainty

The modern information theory is rooted in the work of Shannon (1948). Shannon related the notion of information provided by a probability distribution for predicting outcomes to uncertainty and "choice" at a very basic and intuitive level: A source transmits discrete signals $\mathcal{S} = \{x_1, \cdots, x_n\}$ through a noiseless channel according to a probability distribution $f = (f_1, \cdots, f_n),\ f_i = P(x_i)$. Shannon posed questions such as: "how much information is 'produced'? ... how much 'choice' is involved in the selection of the event or how uncertain we are of the outcome?" (Shannon, 1948, pp. 48-49). He listed a set of properties as reasonable requirements of an uncertainty function $\mathcal{U}(f)$. These properties have been refined and restated in various equivalent forms by Khinchin (1957), Rényi (1961), and others. We state these properties using our notations:

1. *Continuity*: $\mathcal{U}(f_i, 1 - f_i)$ is continuous in f_i.
2. *Symmetry*: $\mathcal{U}(f) = \mathcal{U}(f_1, \cdots, f_n)$ is invariant under permutations of $f_i,\ i = 1, \cdots, n$.
3. *Monotonicity*: $\mathcal{U}(f^*) = \mathcal{U}\left(\frac{1}{n}, \cdots, \frac{1}{n}\right)$, is a monotonically increasing function of $n = 1, 2, \cdots$.

4. *Partition invariance*: If the signals are grouped into disjoint subsets $E_j = \{x_k,\ k = 1, \cdots, n_j\}$ such that $\mathcal{S} = \bigcup_{j=1}^{J} E_j$, then

$$\mathcal{U}[f(x)] = \mathcal{U}(p_1, \cdots, p_J) + \sum_{j=1}^{J} p_j \mathcal{U}[f(x; E_j)], \tag{1}$$

where $p_j = P(E_j) = \sum_{k=1}^{n_j} f(x_k)$, and $f(x; E_j) = \dfrac{f(x)}{P(E_j)}, x \in E_j$ is the truncated distribution.

We refer to the above properties as Shannon's axioms; also see Maasoumi (1993). Rényi (1961) gave an alternative to axiom 4 in terms of invertible functions of $\mathcal{U}(f)$, which reduces to (1) when the function is linear.

Axiom 1 is for mathematical use. The last three axioms have intuitive appeals (Shannon, 1948) and are adaptable to continuous and countable $\mathcal{S}$, as well. By Axiom 2, $\mathcal{U}(f)$ is a function of the magnitudes of probabilities (hight of density), irrespective of the location. Axiom 3 can be represented more generally as follows:

3A. $\mathcal{U}(f^*)$, where $f^*(x) \propto k$, is increasing in the volume (size) of $\mathcal{S}$.

The uniform distribution in 3A is proper, $f^*(x) = \dfrac{1}{||\mathcal{S}||},\ ||\mathcal{S}|| < \infty$, where $||\mathcal{S}||$ denotes the volume of $\mathcal{S}$; otherwise $f^*(x)$ is improper. Axiom 4 can be represented more generally as follows:

4A. For a pair of random variables with joint pdf $f(x_1, x_2)$,

$$\mathcal{U}[f(x_1, x_2)] = \mathcal{U}[f(x_1)] + \int_{\mathcal{S}} \mathcal{U}[f(x_2|x_1)] dF(x_1), \tag{2}$$

where $dF(x) = f(x)dx$ for continuous F, otherwise $dF(x) = f(x)$ and the integral is a sum. For X and $\boldsymbol{Z} = (Z_1, \cdots, Z_J)$ where $Z_j = \phi_j(X)$ is the indicator function of $x \in E_j$, (2) gives (1).

A weaker condition than (2) is the additivity under independence:

4B. $\mathcal{U}[f(x_1)f(x_2)] = \mathcal{U}[f(x_1)] + \mathcal{U}[f(x_2)]$.

An uncertainty function is said to be *additive* if it satisfies 4B. Clearly, (2) implies 4B. Only a few known families of uncertainty measures are additive (Kapur, 1994). The set of additive measures is closed under linear combinations which makes the set expansive through simple constructions, e.g., if $\mathcal{U}_1(f)$ and $\mathcal{U}_2(f)$ are additive, so is $\mathcal{U}(f) = a_1\mathcal{U}_1(f) + a_2\mathcal{U}_2(f)$. Some alternatives to additivity have been proposed in the literature; for example, see the pseudo-additivity defined by (10) in Section 3.2.

A key property of the measures that satisfy various versions of the above axioms is that the maximum uncertainty is attained when the distribution is the uniform. This is in accord with Laplace's "Principle of Insufficient Reason", which in the absence of any information about the outcomes other than $||\mathcal{S}||$, assigns equal probabilities to all possible events of equal size or volumes. This principle implies that the uniform distribution reflects the most unpredictable situation. Any distribution more concentrated than the uniform distribution is more informative for prediction.

Another key property of the unique solution to Shannon's axioms is concavity of $\mathcal{U}(f)$. Like uniformity, concavity is not a defining property of Shannon's measure. However, due to this property "any experiment is informative, on the average" (Lindley, 1956). More generally, an important question in the data analysis is that to what extent the use of a variable X_1 affects uncertainty about predicting the outcomes of another variable X_2. DeGroot (1962) asserted that the worst case scenario is when the outcomes of one variable, on average, have no effect on uncertainty about

prediction of another variable. He showed that this assertion holds if and only if $\mathcal{U}(f)$ is concave in f. Concavity is also necessary for the solutions to derivation of probability distributions based on partial information by the maximum entropy method (Jaynes, 1957) to be unique, when they exist.

The foregoing summary of literature suggests that two desirable properties for an uncertainty function of a probability distribution F are:

(a) $\mathcal{U}(f)$ is a concave scalar function of f;
(b) $\mathcal{U}(f) \leq \mathcal{U}(f^*)$, where f^* is the uniform,

where f is a pdf relative to a dominating measure.

The property (b), referred to as the uniformity requirement, is a modification of the definition of uncertainty function of Goel and DeGroot (1981) where concavity is the only requirement. For example, concavity includes variance (Goel and DeGroot, 1981), but variance is not a general measure of uncertainty. It applies only when the outcomes are quantitative and the distribution is univariate. The natural extension of variance to the multivariate case is the dispersion matrix which cannot be summarized uniquely in terms of a scalar function of f (see Stone, 1959). Under certain conditions, the variance maps uniformity, but this is not universally true (Ebrahimi, et al., 1999). For some distributions, variance is not defined and for some others it does not map the lack of concentration of probabilities. Figure 1 illustrates all these cases. The uncertainty can be compared by variance for the two Gaussian distributions shown in (a). The variance of Cauchy distribution shown in (b) is not defined. The exponential and gamma distributions shown in (c) have the same variance, but clearly display unequal levels of concentrations, hence unequal levels of difficulty of predictability. The beta distribution shown in (d) is more concentrated but has a larger variance than the uniform distribution. In this case, variance contradicts lack of concentration of the distribution. As will be seen in Section 3, a measure $\mathcal{U}$ having the uniformity property orders each pair of the distributions shown in Figure 1.

As a final remark, the distribution F itself can be subject to uncertainty, fully or partially, e.g., $F = F_\theta$, where θ is an unknown parameter. Then $\mathcal{U}(f)$ is also subject to uncertainty. The uncertainty about F can be mapped by a distribution Φ with its support being a set of distributions $\Omega_F = \{F\}$. Estimates of $\mathcal{U}(f)$ can be found, for example, by the expected values $\tilde{\mathcal{U}}(f) = E_\Phi[\mathcal{U}(f)]$.

2.2 Notion of Information

Broadly speaking, information refers to the changes that knowledge induces to the probability distribution used for inference. In the absence of information, the uniform distribution f^* is used and thus it is the *global reference distribution* for quantifying information in terms of unpredictability. The information provided by a distribution f for prediction of outcomes is quantified by its discrepancy with the uniform distribution f^*. An example of such information discrepancy measure is the uncertainty difference,

$$\mathcal{D}(f : f^*) = \Delta\mathcal{U}(f : f^*) = \mathcal{U}(f^*) - \mathcal{U}(f) \geq 0, \tag{3}$$

where the inequality is implied by the uniformity property of $\mathcal{U}$. By properties (a) and (b), the equality holds if and only if $f(x) = f^*(x)$ almost everywhere. The information discrepancy $\mathcal{D}(f : f^*)$ is convex in f if and only if the uncertainty function $\mathcal{U}(f)$ is concave in f.

Kullback and Leibler (1951) generalized Shannon's notion of information in terms of discrepancy between two distributions beyond the finite discrete and Lebesgue measures. They did not begin

with a set of postulates but showed that their measure has some desirable properties, which were further explored by Kullback (1954). Among the properties of the information discrepancy are convexity and non-negativity (Kullback, 1959); see also Burbea and Rao (1982). Following this line, we conceptualize information provided by f_1 relative to a reference distribution f_2 as an information discrepancy function that has the following two desirable properties:

(a) $\mathcal{D}(f_1 : f_2) \geq 0$, where the equality holds if and only if $f_1(x) = f_2(x)$ almost everywhere;
(b) Given f_2, $\mathcal{D}(f_1 : f_2)$ is convex in f_1.

In general, $\mathcal{D}(f_1 : f_2)$ does not indicate which of the two distributions is more informative for prediction. When the reference distribution is uniform, $f_2 = f^*$, then $\mathcal{D}(f_1 : f^*)$ quantifies the information provided by f_1 for prediction.

3 Measures of Uncertainty

This section presents Shannon entropy and two of its generalizations.

3.1 Shannon Entropy

Shannon entropy of a distribution with pdf $f(x)$ is defined by

$$H(X) \equiv H(f) = -\int_{\mathcal{S}} \log f(x) dF(x). \tag{4}$$

For $\mathcal{S} = \{x_1, \cdots, x_n\}$, the entropy is the unique solution to Axioms 1-4. Rényi (1961) showed the same result for an incomplete distribution, $0 < \sum_{i=1}^{n} f(x_i) \leq 1$. For general $\mathcal{S}$, the entropy satisfies Axiom 4A, which is the defining property of this measure. The entropy is concave and $H(f) \leq \log ||\mathcal{S}||$, where the equality holds if and only if f is uniform (improper when $\mathcal{S}$ is unbounded).

Example 3.1 The distributions shown in Figure 1 are ordered by the entropy as follows:

Distribution:	Beta	Uniform	Gamma	Exponential	Gaussian	Gaussian	Cauchy
Parameters:	(.5,.5)	(0,1)	(1,1.41)	1	(0,1)	(0,4)	(0,1)
Entropy:	-.242	0	.924	1	1.419	2.112	2.531

Entropy expressions for univariate and multivariate families of distributions are available, e.g., in Cover and Thomas (1991) and Nadarajah and Zografos (2005).

The *conditional entropy* of X_1 given X_2 is defined by

$$\mathcal{H}(X_i|X_j) = \int_{\mathcal{S}_j} H(X_i|x_j) dF(x_j) \leq H(X_i), \quad i \neq j = 1, 2, \tag{5}$$

where the inequality is implied by concavity of $H(f)$ and becomes equality if and only if X_1 and X_2 are independent. The script $\mathcal{H}$ is used to emphasize that $\mathcal{H}(X_i|X_j)$ is the average of the entropies of all conditional distributions $f(x_i|x_j)$ with respect to the distribution of X_j. By (2) and (5), the entropy is sub-additive, i.e.,

$$H(X_1, X_2) = H(X_1) + \mathcal{H}(X_1|X_2) \leq H(X_1) + H(X_2), \tag{6}$$

where the equality holds if and only if X_1 and X_2 are independent.

There are some notable differences between entropies in the finite and general cases of $\mathcal{S}$. In the finite case, $H(X) \geq 0$, where the equality holds if and only if f is a degenerate distribution. In general, $-\infty \leq H(X) \leq \infty$ and $H(f) = 0$ does not imply F is degenerate. Bound for the entropy in terms of moments are presented in Section 5.

In the discrete case, in general, transformation decreases entropy (Cover & Thomas, 1991, p. 43), and $H(X)$ is invariant under one-to-one transformations of X. The continuous entropy is not invariant under all one-to-one transformations of X. Let $\phi : \Re^d \to \Re^d$ be one-to-one and $\boldsymbol{Y} = \phi(\boldsymbol{X})$. Then

$$H(\boldsymbol{Y}) = H(\boldsymbol{X}) + E[\log J_\phi(\boldsymbol{Y})], \tag{7}$$

where $J_\phi(\boldsymbol{Y}) = \left|\left[\frac{\partial \phi^{-1}(y_i)}{\partial y_k}\right]\right|$, $i, k = 1, \cdots, d$ is the Jacobian of transformation; see, e.g., Darbellay and Vajda (2000). Thus, for the continuous case, $H(\boldsymbol{Y})$ can be larger or smaller than, or equal to $H(\boldsymbol{X})$ depending on $E[\log J_\phi(\boldsymbol{Y})]$. By (7), $H(\boldsymbol{X})$ is invariant under translation and under orthonormal transformation, $H(A\boldsymbol{X}) = H(\boldsymbol{X})$, where A is $d \times d$ matrix with determinant $|A| = 1$.

A useful representation of entropy for the univariate X is the quantile formula given by

$$H(f) = -\int_0^1 \log\left[\frac{dQ(u)}{du}\right]^{-1} du,$$

where $Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\}$ is the quantile function. This representation, first noted by Vasicek (1976), has played an important role for developing inferences about entropy and information indices of fit. The quantile representation facilitates approximation of the entropy when F is absolutely continuous. Consider the quantized entropy

$$H_{m,q}(F) = -\sum_{k=1}^{q} \Delta F_{0,k} \log \frac{\Delta F_{m,k}}{\Delta \xi_{m,k}}, \tag{8}$$

where $-\infty < \xi_1 < \cdots < \xi_q \leq \infty$, $\Delta\xi_{m,k} = \xi_{k+m} - \xi_{k-m}$ is spacing of order $2m \in \{0, 2, 4, \cdots, 2q\}$, $\Delta F_{m,k} = F_{k+m} - F_{k-m}$, $\Delta\xi_{0,k} \equiv \xi_k - \xi_{k-1}$, $\Delta F_{0,k} \equiv F(\xi_k) - F(\xi_{k-1})$, $\xi_{k-m} \equiv \xi_0 = \sup\{x : F(x) = 0\} \geq -\infty$ for $k < m$, and $\xi_{k+m} \equiv \xi_q$ for $k > q - m$. For suitably chosen ξ_i's, $H_{m,q}(F) \xrightarrow{p} H(X)$. For $m = 0$, (8) is a modification of the quantized approximation of the continuous entropy by the discrete entropy (Cover and Thomas, 1991). The quantized entropy includes several entropy estimates proposed in the literature; see Mazzuchi, et al. (2008) for details and references.

A useful representation of entropy for reliability analysis is

$$H(f) = 1 - \int_0^\infty f(x) \log \lambda(x) dF(x), \tag{9}$$

where $\lambda(x) = \frac{f(x)}{\bar{F}(x)}$ is the failure (hazard) rate (Teitler, et al., 1986) and $\bar{F}(x) = 1 - F(x)$ denotes the survival function. The hazard rate representation has led to some results in reliability analysis. For example, (9) is used in Ebrahimi (1996) for developing several results.

3.2 Generalizations

The entropy of order α of a distribution (Rényi, 1961) is defined as

$$H_{R,\alpha}(X) \equiv H_{R,\alpha}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{S}} f^\alpha(x) d\nu(x),$$

where $d\nu(x) = dx$ for the continuous, $d\nu(x) = 1$ for discrete cases, and $\alpha > 0,\ \alpha \neq 1$. It is well known that $H_{R,1}(f) = \lim_{\alpha\to 1} H_{R,\alpha}(f) = H(f)$. Rényi entropy expressions for many univariate, bivariate, and multivariate distributions are given in Nadarajah and Zografos (2003, 2005) and Zografos and Nadarajah (2005).

Rényi (1961) showed that for $\mathcal{S} = \{x_1, \cdots, x_n\}$, $H_{R,\alpha}(X)$ satisfies Shannon's Axioms 1-3 and a weaker condition than Axiom 4 which he formulated as an alternative. However, $H_{R,\alpha}(X)$ is additive for independent random variables. For $\alpha \leq 1$, $H_{R,\alpha}(X)$ is concave when $f(x) \leq 1, \forall x \in \mathcal{S}$, e.g., the discrete case. Otherwise, $H_{R,\alpha}(X)$ is neither concave, nor convex. The conditional Rényi entropy $\mathcal{H}_{R,\alpha}(X_1|X_2)$ is defined similarly as in (5), where the inequality holds only for $\alpha \leq 1$ and $f(x) \leq 1, \forall x \in \mathcal{S}$. Rényi entropy satisfies a weaker property than (5) formulated by Jizba and Arimitsu (2004) in terms of conditioning on events. Moreover, there is no useful transformation formula like (7) for $H_{R,\alpha}(X)$.

Tsallis (1988) defined an entropy measure which can be represented as

$$\begin{aligned} H_{T,\alpha}(X) \equiv H_{T,\alpha}(f) &= -\int_{\mathcal{S}} f^{\alpha}(x) L_{\alpha}\left(f(x)\right) d\nu(x) \\ &= \frac{1}{1-\alpha}\int_{\mathcal{S}} \left[f^{\alpha}(x) - 1\right] d\nu(x), \end{aligned}$$

where $d\nu(x)$ is defined as above, and

$$L_{\alpha}(z) = \begin{cases} \dfrac{z^{1-\alpha}-1}{\alpha-1}, & \alpha \neq 1, \\ \log z, & \alpha = 1 \end{cases}$$

is referred to as a generalized logarithm function and $\lim_{\alpha\to 1} L_{\alpha}(z) = \log z$. $H_{T,\alpha}(f)$ is also known as Tsallis-Havrda-Charvat (THC) entropy and $L_{\alpha}(\cdot)$ is pseudo-additive defined as:

$$L_{\alpha}(x_1, x_2) = L_{\alpha}(x_1) + L_{\alpha}(x_2) + (1-\alpha) L_{\alpha}(x_1) L_{\alpha}(x_2). \tag{10}$$

This property directly applies to THC for independent random variables X_1 and X_2, where $L_{\alpha}(\cdot)$ is replaced with $H_{T,\alpha}(\cdot)$ in (10). Like Shannon entropy, $H_{T,\alpha}(f)$ is concave, so (5) holds, where the conditional THC entropy $\mathcal{H}_{T,\alpha}(X_1|x_2)$ is defined similarly.

A list of several generalizations of Shannon entropy is given in Esteban and Morales (1995).

4 Measures of Information

This section presents KL information, its symmetric versions, and three of its generalizations.

4.1 KL Information

The KL discrimination information between two probability distributions $F_k,\ k = 1, 2$ with pdfs f_k is defined by

$$K(f_1 : f_2) = \int_{\mathcal{S}} \log \frac{f_1(x)}{f_2(x)} dF_1(x), \tag{11}$$

provided that F_1 is absolutely continuous with respect to F_2, denoted as $F_1 \ll F_2$. This condition is necessary, but not sufficient, for finiteness of $K(f_1 : f_2)$; see Example 4.1. The term *information*, in addition to its historical background mentioned in Section 2, is reflective of the facts that $K(f_1 : f_2)$ is a generalization of (4) and that $K(f_1 : f_2)$ is the expected log-odds in favor of F_1

given by the Bayes rule (Kullback, 1959). Rényi (1961) showed that for $\mathcal{S} = \{x_1, \cdots, x_n\}$, $K(f_1 : f_2)$ uniquely satisfies a set of properties which are analogous to Axioms 1-4. (Rényi considered incomplete distributions and included another axiom which for complete probability distributions $\sum_{i=1}^{n} f_k(x_i) = 1,\ k = 1, 2$ is equivalent to the non-negativity of information discrepancy).

The relationship (3) holds for Shannon entropy and the KL information:

$$\mathcal{D}(f : f^*) = K(f : f^*) = H(f^*) - H(f).$$

The quantity

$$I(f) = -H(f) = E_f[\log f(X)] \tag{12}$$

is the average log-height of the pdf f and is referred to as Shannon information in statistics literature (Lindley, 1956, Zellner, 1971). A useful representation of KL is

$$K(f_1 : f_2) = H_{f_1}(f_2) - H(f_1) = I(f_1) - I_{f_1}(f_2), \tag{13}$$

where

$$I_g(f) = E_g([\log f(X)] \tag{14}$$

is known as Fraser information. This measure is motivated by Fraser (1965) and used by Kent (1982, 1983) and others in terms of the "information gain" about a parameter, $I_g[f(x|\theta)] = -H_g[f(x|\theta)]$, where $f(x|\theta)$ is a parametric model and $g(x)$ is "true" density. Clearly, $I_f(f) = I(f)$. Note that $I(f)$ and $I_g(f)$ are convex in f, but can be negative.

Properties of $K(f_1 : f_2)$ are the same for distributions with all types of $\mathcal{S}$. We will present a few properties. See Kullback (1959), Soofi and Retzer (2002), and Ebrahimi and Soofi (2004) for other properties and more interpretations. $K(f_1 : f_2) \geq 0$, where equality holds if and only if $f_1(x) = f_2(x)$ almost everywhere. It is an information discrepancy, also referred to as cross-entropy, relative entropy, and directed divergence between f_1 and f_2, but it is not a distance function (see Kullback, 1987). For example, $K(f_1 : f_2)$ is not symmetric.

For two random variables X_1 and X_2, the additive decomposition is

$$K[f_1(x_1, x_2) : f_2(x_1, x_2)] = K[f_1(x_1) : f_2(x_1)] + E_1\left\{K[f_1(x_2|x_1) : f_2(x_2|x_1)]\right\}. \tag{15}$$

Thus, (11) is additive for independent random variables.

In general, transformation reduces information. Let $\phi(\cdot)$ be a function, $W = \phi(X)$ and $g_k(w)$ denote the pdf induced by $f_k(x),\ k = 1, 2$. Then $K(g_1 : g_2) \leq K(f_1 : f_2)$, where the equality holds if and only if $\phi(\cdot)$ is sufficient for discrimination: $\dfrac{g_1(y)}{g_2(y)} = \dfrac{f_1(x)}{f_2(x)}$, almost everywhere. Examples includes sufficient statistics when $f_k = f_k(x|\theta)$ and when ϕ is a one-to-one transformation of X.

Let $\boldsymbol{Z} = \phi(X)$ be a vector of indicator functions $\boldsymbol{Z} = (Z_1, \cdots, Z_n)$ of the partition of $\mathcal{S}$ and $\boldsymbol{p}_k = (p_{k1}, \cdots, p_{kn}),\ p_{ki} = P_k(E_i) = \int_{E_i} dF_k(x) > 0,\ k = 1, 2$. Application of (15) gives

$$K(f_1 : f_2) = K(\boldsymbol{p_1} : \boldsymbol{p_2}) + \sum_{i=1}^{n} P_1(E_i) K(f_1 : f_2; E_i) \geq K(\boldsymbol{p_1} : \boldsymbol{p_2}), \tag{16}$$

where $K(f_1 : f_2; E_i)$ is the discrimination information between the truncated distributions $f_k(x; E_i) = \dfrac{f_k(x)}{P_k(E_i)}$. The inequality in (16) illustrates that grouping leads to loss of information unless it is sufficient for discrimination: $\dfrac{f_1(x)}{f_2(x)} = \dfrac{P_1(E_i)}{P_2(E_i)}$ for all $x \in E_i$ and for all $i = 1, \cdots n$. When $n = 2$, the

inequality in (16) gives the calibration measure proposed by McCulloch (1989) which is interpreted in terms of discrimination information between the flips of a fair coin and a biased coin.

It should be noted that the lack of symmetry is not of a concern if either F_1 or F_2 is an ideal distribution (e.g., the true data-generating distribution) or an initial distribution to be updated in light of data. Then $K(f_1 : f_2)$ measures loss of information in using the other distribution instead of the ideal one and gain of information by use of the enhanced distribution; see Section 5. However, in some applications, the lack of symmetry can be an issue.

A symmetric version, referred to as Jeffreys divergence (Jeffreys 1946) measure, is given by

$$J(f_1, f_2) = K(f_1 : f_2) + K(f_2 : f_1).$$

This measure requires $F_1 \ll F_2$ and $F_2 \ll F_1$, which is more stringent than the requirement for $K(f_1 : f_2)$. Bernardo and Rueda (2002) defined the *intrinsic information* measure as

$$\delta(f_1, f_2) = \min\{K(f_1 : f_2), K(f_2 : f_1)\}.$$

This symmetric measure bypasses the absolute continuity requirement. If $F_1 \ll F_2$ does not hold, we have $K(f_1 : f_2) = \infty$ and $\delta(f_1, f_2) = K(f_2 : f_1)$. The following example illustrates these information measures.

Example 4.1 Let f_s and f_d be the pdfs shown as solid and dashed curves in Figure 1, respectively. The information discrepancy measures for each pair are as follows:

Panel:	(a)	(b)	(c)	(d)
$K(f_s : f_d)$:	1.011	(3.103, 3.603)	.645	.145
$K(f_d : f_s)$:	.807	∞	.530	.242
$J(f_s, f_d)$:	1.818	∞	.387	1.175
$\delta(f_s, f_d)$:	.807	(3.103, 3.603)	.530	.145

Two KL measures are computed for distributions in Panels (a), (c), and (d). For Panel (b), $K(f_s : f_d) = K_0 + \frac{1}{2}E_{f_s}[\log(1 + X^2)]$, where $K_0 = \frac{3}{2}\log\pi + \frac{1}{2}\log 2 = 3.103$. Noting that $0 < \log(1 + x^2) \le x^2$, we have $K_0 < K(f_s : f_d) \le K_0 + \frac{1}{2}E_{f_s}(X^2) = K_0 + \frac{1}{2}$. The Cauchy and normal distributions are absolutely continuous with respect to each other, but $K(f_d : f_s)$ is not finite. Computation of $K(f_d : f_s)$ requires $E_{f_d}(X^2)$ which is not finite. Consequently, Jeffreys divergence measure cannot be computed for Panel (b). The intrinsic information measures are computed for all four panels. The case of Panel (b) illustrates usefulness of $\delta(f_1, f_2)$ when the absolute continuity holds in both directions.

4.2 Generalizations

Rényi (1961) information divergence of order α between two distributions is defined by

$$K_{R,\alpha}(f_1 : f_2) = \frac{1}{\alpha - 1}\log\int_{\mathcal{S}} f_1^{\alpha}(x) f_2^{1-\alpha}(x) d\nu(x), \quad \alpha \neq 1.$$

It is well known that $K_{R,1}(f_1 : f_2) = \lim_{\alpha\to 1} K_{R,\alpha}(f_1 : f_2) = K(f_1 : f_2)$. Only for $\alpha \ge 1$ the absolute continuity $F_1 \ll F_2$ is needed. Like (11), $K_{R,\alpha}(f_1 : f_2)$ is nonnegative, invariant under one-to-one transformations of X, and additive for independent random variables X_1 and X_2, but general additive decomposition (15) does not hold for $K_{R,\alpha}(f_1 : f_2)$.

Tsallis (1998) defined a generalization of KL information for probability vectors which can be represented more generally as

$$\begin{aligned} K_{T,\alpha}(f_1 : f_2) &= \int_{\mathcal{S}} f_1(x) L_\alpha \left(\frac{f_1(x)}{f_2(x)} \right) d\nu(x) \geq 0 \\ &= \frac{1}{\alpha - 1} \int_{\mathcal{S}} \left[f_1^\alpha(x) f_2^{1-\alpha}(x) - 1 \right] d\nu(x), \quad \alpha \neq 1. \end{aligned} \tag{17}$$

The equality holds if and only if $f_1(x) = f_2(x)$ almost everywhere, and $K_{T,1}(f_1 : f_2) = \lim_{\alpha \to 1} K_{T,\alpha}(f_1 : f_2) = K(f_1 : f_2)$. Like (11), $K_{T,\alpha}(f_1 : f_2)$ is nonnegative and invariant under one-to-one transformations of X. For two independent X_1 and X_2, (17) is pseudo-additive as defined in (10).

Cressie and Read (1984) defined a measure of divergence for multinomial distributions which can be represented as

$$K_{C,\alpha}(f_1 : f_2) = \frac{1}{\alpha} K_{T,\alpha}(f_1 : f_2), \quad \alpha \neq 0, 1. \tag{18}$$

It is well known that $\lim_{\alpha \to 0} K_{C,\alpha}(f_1 : f_2) = \lim_{\alpha \to 1} K_{C,\alpha}(f_1 : f_2) = K(f_1 : f_2)$. Cressie and Read (1984) used the discrete version of this measure for multinomial estimation. Imben, et al. (1998) proposed an information theoretic approach to the generalized method of moments based on (18).

The case of $\alpha = \frac{1}{2}$ is of particular interest since $K_{R,1/2}(f_1 : f_2)$, $K_{T,1/2}(f_1 : f_2)$, and $K_{C,1/2}(f_1 : f_2)$ are all symmetric in f_1 and f_2. Furthermore,

$$\begin{aligned} K_{T,1/2}(f_1 : f_2) &= K_{T,1/2}(f_2 : f_1) \\ &= \int_{\mathcal{S}} \left[f_1^{1/2}(x) - f_2^{1/2}(x) \right]^2 d\nu(x) \\ &= 2 \left[1 - \int_{\mathcal{S}} f_1^{1/2}(x) f_2^{1/2}(x) d\nu(x) \right]. \end{aligned} \tag{19}$$

The first integral is a Hellinger distance and the second integral is known as Bhattacharya distance.

5 Information Criteria

Maximum entropy information criteria (MEIC) develops probability models which are most noncommittal to information other than that explicitly taken into account. The MEIC extends Laplace's principle of insufficient reason. The minimum discrimination information criteria (MDIC), also known as the minimum cross-entropy principle, generalizes the MEIC by developing noncommittal models with reference to any given measure instead of the uniform reference.

5.1 MDIC and MEIC

Given f_2, $K(f_1 : f_2)$ is convex in f_1. Therefore, for a class of distributions $K(f_1 : f_2)$ can be minimized with respect to f_1. Consider the moment class of distributions:

$$\Omega_{\boldsymbol{\theta}} = \{ f(x|\boldsymbol{\theta}) : E_f[T_j(X)|\boldsymbol{\theta}] = \theta_j, \ j = 1, \cdots, J \}, \tag{20}$$

where $T_j(X)$ are integrable with respect to $dF(x)$ and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_J)$.

Definition 1 *The MDI model in $\Omega_{\boldsymbol{\theta}}$ reference to f_0 is $f^* = \arg\min_{f \in \Omega_{\boldsymbol{\theta}}} K(f : f_0)$.*

The MDI model $f^* \in \Omega_{\boldsymbol{\theta}}$ is unique and is in the form of

$$f^*(x, f_0|\boldsymbol{\theta}) = C_0(\boldsymbol{\lambda}) f_0(x) e^{-\lambda_1 T_1(x) - \cdots - \lambda_J T_J(x)}, \tag{21}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_J)$, $\theta_j = \dfrac{\partial}{\partial \lambda_j} \log C_0(\lambda)$, is the vector of Lagrange multipliers, provided that the normalizing factor $C_0(\boldsymbol{\lambda})$ is strictly positive and finite. For all $f \in \Omega_{\boldsymbol{\theta}}$,

$$K(f : f_0|\boldsymbol{\theta}) \geq K(f^* : f_0|\boldsymbol{\theta}) = \log C_0(\boldsymbol{\lambda}) - \boldsymbol{\lambda}'\boldsymbol{\theta}. \tag{22}$$

These results also hold when $\Omega_{\boldsymbol{\theta}}$ in (20) is defined by $E_f[T_j(X)|\boldsymbol{\theta}] \geq \theta_j$; see Appendix in Kullback (1959, reprinted in 1968) and Shore and Johnson (1980).

The MEIC proposed by Jaynes (1957, 1968) gives the following criterion.

Definition 2 *The ME model in $\Omega_{\boldsymbol{\theta}}$ defined in (20) is $f^* = \arg\max_{f \in \Omega_{\boldsymbol{\theta}}} H[f(x)]$.*

When f_0 in (21) is uniform (proper or improper), the MDIC and MEIC are equivalent.

If (21) exists when f_0 is a constant, then (21) gives the unique ME model

$$f^*(x|\boldsymbol{\theta}) = C(\boldsymbol{\lambda}) e^{-\lambda_1 T_1(x) - \cdots - \lambda_J T_J(x)}, \tag{23}$$

and for all $f \in \Omega_{\boldsymbol{\theta}}$,

$$H[f(X|\boldsymbol{\theta})] \leq H[f^*(x|\boldsymbol{\theta})] = -\log C(\boldsymbol{\lambda}) + \boldsymbol{\lambda}'\boldsymbol{\theta}. \tag{24}$$

In particular, when $\Omega_\theta = \left\{ f(x|\theta) : E[X|^k = \theta \right\}$, (24) gives the entropy-moment inequality:

$$H(X) \leq \frac{1}{k} \log \frac{2^k e \Gamma^k (1/k) E|X|^k}{k^{k-1}}, \tag{25}$$

where the equality is attained by the ME model in Ω_θ. The well-known examples are the exponential distribution for $k = 1,\ x \geq 0$, double-exponential (Laplace) distribution for $k = 1,\ x \in \Re$ and normal (Gaussian) distribution for $k = 2$. By (25), the entropy of distributions with a finite variance is finite. But the converse is not true. Bound for the entropy of discrete distribution in terms of the entropy of a continuous distribution with given variance is presented in Cover and Thomas, 1991, pp. 235-236). By (24) and (25), their procedure can be used to obtain bounds for the entropy of distributions based on various moments.

For any $f \in \Omega_{\boldsymbol{\theta}}$ and $f^* \in \Omega_{\boldsymbol{\theta}}$, we have the information distinguishability (ID) relationship

$$K(f : f^*|\boldsymbol{\theta}) = H(f^*|\boldsymbol{\theta}) - H(f|\boldsymbol{\theta}), \tag{26}$$

where $H(f^*|\boldsymbol{\theta}) = H(f|\boldsymbol{\theta})$ if and only if $f(x|\boldsymbol{\theta}) = f^*(x|\boldsymbol{\theta})$ almost everywhere (Soofi, et al. 1995, Ebrahimi, et al. 2008). Comparison of (26) with (13) reveals that for $f, f^* \in \Omega_{\boldsymbol{\theta}}$, Fraser information is negative Shannon entropy, $I_f(f^*) = I(f^*)$. The ID relationship (26) is a simple but sufficiently general result which plays the key role in some information methods. Next, we briefly mention a few. Application to model selection will be presented in Section 5.4.

Application of (26) simplifies the ME characterization problem to identifying the information moment set by (23). Ebrahimi, et al. (2008) showed that any distribution with a pdf in the form of (23) is the unique ME model in the moment class of distributions (20) generated by the information moment set $\mathcal{T}_X = \{T_j(X), j = 1, \cdots, J\}$ shown in the exponent of the pdf in (23). For a parametric model, one may easily identify the moment class $\Omega_{\boldsymbol{\theta}}$ by writing the pdf in the exponential form

(23). Many known univariate and multivariate parametric families of distributions are in the form of (23) and therefore are ME subject to specific forms of moment constraints.

The ID relationship (26) also facilitates derivations of limiting distributions. Barron (1986) proved the Central Limit Theorem (CLT) for given variance. In the CLT case (26) is applicable and convergence in entropy is equivalent with $K(f_1 : f_2) \to 0$. Convergence in discrimination information $K(f_1 : f_2) \to 0$ implies convergence in distribution due to the results available based on L_1-norm and the inequality $||f_1 - f_2||^2 \leq 2K(f_1 : f_2)$, see Barron (1986) for references. In problems such as convergence to an extreme value distribution, the limiting distribution is not the ME model in the class of models that contains the sequence $\Omega = \{F_n\}$, thus (26) does not apply, and convergence in entropy does not imply $K(f_1 : f_2) \to 0$.

Successive application of (26) to the moment constraints in (20) gives

$$K(f : f^*|\boldsymbol{\theta}) = \sum_{j=1}^{J} K(f^*|\boldsymbol{\theta}_{j-1}, f^*|\boldsymbol{\theta}_j) = \sum_{j=1}^{J} \Delta H(f^*|\boldsymbol{\theta}_j, f^*|\boldsymbol{\theta}_{j-1}), \tag{27}$$

where $f^*(x|\boldsymbol{\theta}_0) = f(x|\boldsymbol{\theta})$ and $f^*(x|\boldsymbol{\theta}_j) = f(x|\theta_1, \cdots, \theta_j)$, provided that their entropies are finite. Applications of (27) include the "analysis of information" for categorical data (Gokhale and Kullback, 1978) and information indices of predictors in logit models defined by the normalized information differences (Soofi, 1992, 1994).

Example 5.1 (Location-scale family) A pdf $f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is said to be in the multivariate location-scale family of distributions Ω_{LS} with location vector $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$ if

$$f(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}_d) = |\boldsymbol{\Sigma}|^{1/2} f\left(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{28}$$

where $\boldsymbol{0}$ is the vector of zeros and I_d is the identity matrix. The best known example in the family (28) is the multivariate normal distribution with pdf

$$f^*(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}, \quad \boldsymbol{x} \in \Re^d.$$

This pdf is in the form of (23) with $T_1(\boldsymbol{X}) = \boldsymbol{X}$ and $T_2(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'$. Thus, $f^*(\boldsymbol{x})$ is ME in $\Omega_{\boldsymbol{\theta}} = \{f : E_f = \boldsymbol{\mu}, E_f[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}\}$. Note that $\Omega_{\boldsymbol{\theta}} \subset \Omega_{LS}$, since mean and covariance for some $f \in \Omega_{LS}$ are not defined (e.g., Cauchy). For any $f \in \Omega_{LS}$ with a finite entropy,

$$H(\boldsymbol{X}; \boldsymbol{\Sigma}) = H(\boldsymbol{X}; \boldsymbol{I}_d) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \leq H(\boldsymbol{X}^*; \boldsymbol{\Sigma}),$$

where $H(\boldsymbol{X}; \boldsymbol{I}_d)$ is a function of parameters of $f(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I}_d)$ and the dimension d and

$$H(\boldsymbol{X}^*; \boldsymbol{\Sigma}) = H(f^*) = \frac{d}{2} + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| \tag{29}$$

is the normal entropy. If $\Omega_{\boldsymbol{\theta}_0} \subset \Omega_{LS}$ where, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = Diag[\sigma_1^2, \cdots, \sigma_d^2]$, then the ME model $f_0^*(\boldsymbol{x}) \in \Omega_{\boldsymbol{\theta}_0}$ is the independent normal distribution. Since $\Omega_{\boldsymbol{\theta}} \subseteq \Omega_{\boldsymbol{\theta}_0}$, the ID relationship (26) is applicable and the entropy reduction due to the additional off-diagonal constraints in $\Omega_{\boldsymbol{\theta}}$ is

$$g(\boldsymbol{R}) = K(f^* : f_0^*) = H(\boldsymbol{X}^*; \boldsymbol{\Sigma}_0) - H(\boldsymbol{X}^*; \boldsymbol{\Sigma}) = -\frac{1}{2}\log|\boldsymbol{R}| \geq 0,$$

where $\boldsymbol{R}$ is the correlation matrix of f^*. Thus, $g(\boldsymbol{R})$ is an example of the ID relationship (26).

5.2 Geometric Interpretation

The MDIC (MEIC) seeks the distribution that is closest to the reference distribution f_0 (uniform distribution) and satisfies the moment conditions that define $\Omega_{\boldsymbol{\theta}}$. A geometric interpretation of the MDI model as the information projection of f_0 on $\Omega_{\boldsymbol{\theta}}$ is given by the following Pythagorean type equality. If $\Omega_{\boldsymbol{\theta}} \subseteq \Omega_0$ where Ω_0 is a linear set of probability vectors, then

$$K(f : f_0) = K(f : f^*) + K(f^* : f_0); \tag{30}$$

Csiszar (1975). Later, Csiszar (1991) developed an axiomatic foundation for the geometric interpretation of MDIC and MEIC for probability vectors, analogous to the Euclidean geometry of the least squares. The relationship (30) is the basis of the MDI tests for information analysis of categorical data (Gokhale and Kullback, 1978) and the information indices of logit models.

A geometric interpretation analogous to (30) can be obtained when $\Omega_{\boldsymbol{\theta}}$ is not a set of probability vectors. If there exists a pdf $f_0 \notin \Omega_{\boldsymbol{\theta}}$ with finite entropy such that $H(f^*|\boldsymbol{\theta}) \leq H(f_0)$, then

$$K[f(x|\boldsymbol{\theta}) : f_0(x)] = K[f(x|\boldsymbol{\theta}) : f^*(x|\boldsymbol{\theta})] + K[f^*(x|\boldsymbol{\theta}) : f_0(x)]. \tag{31}$$

This relationship is obtained by noting that f_0 is the ME model in $\Omega_0 = \Omega_{\boldsymbol{\theta}} \cup \{f_0\}$ and applying (26) to (31). Examples include: (a) $\Omega_\theta = \{f(x|\theta), 0 < x < 1 : E(\log X) = \theta\}$ and $f_0(x) = 1$; (b) $\Omega_{\boldsymbol{\theta}} = \{f(x|\boldsymbol{\theta}), x > 0 : E(X) = \theta_1, E(\log X) = \theta_2\}$ and f_0 is an exponential pdf with mean greater than or equal to θ_1; and (c) $\Omega_\theta = \left\{f(x|\theta), -\infty < x < \infty : E(X^2) = \theta\right\}$ and f_0 is a pdf with $H(f_0) > .5\log(2\pi e\theta)$. Statistical applications of (31) include explication of the projection and decomposition developed by Hastie (1987) for the estimates of KL information measures in terms of the likelihood quantities of the exponential family regression; for the normal model it gives the least squares projection and the sums of squares decomposition of the linear models.

5.3 Axiomatic foundation

Shore and Johnson (1980) provided an axiomatic foundation for the MDIC and MEIC in a general inductive inference framework. In this framework, $f_0(x)$ is viewed as the distribution that reflects the state of knowledge for inference about X before the new information in terms of the moment constraints in (20) was available. They formulated four "consistency" axioms for updating f_0 in light of the new information and showed that the optimal learning model is equivalent to the MDIC. They also showed that MEIC is the optimal learning model for the discrete case when f_0 is uniform. The mathematical formulation of the axioms and the proofs are rather involved. The axioms, stated informally are as follows:

1. *Uniqueness*: The inferential distribution should be unique.
2. *Invariance*: The inferential distribution should not depend on the choice of coordinate system.
3. *System Independence*: The inference based on different densities obtained separately from independent information about independent systems should be the same as that obtained jointly in terms of a joint density.
4. *Subset Independence*: The inference based on treating an independent subset of system states in terms of a separate conditional density should be the same as that in terms of the full system density.

The underpinning principle of these axioms is that "if a problem can be solved in more than one way, the result should be consistent". Thus, the MDIC and MEIC "are not only well behaved in a statistical sense but prove to be inferentially sound" (Gabriele, 1999). The proof is deep and

"rests primarily on the subset independence property" (Shore and Johnson, 1980). Without this property, a generalization in terms of $\int [f(x)/f_0(x)]^\alpha \, dF(x)$ can be obtained which gives MDIC for $\alpha = 1$ (Karbelkar, 1986), but for $\alpha \neq 1$ there is no known explicit solution such as (21).

The MDIC and MEIC are learning models in statistics when $\theta_j,\ j = 1, \cdots, J$ are computed from the data. Examples include the internal constraints problems (Gokhale and Kullback, 1978), ME and MDI logit models (Soofi, 1992, 1994, Golan, et al., 1996), ME model for the regression function (Ryu, 1993), the Bayesian method of moments (Zellner, 1996, 1997) where post data distributions are developed for parameters and prediction, and ME model fitting based on (26) discussed next.

5.4 Model Selection

In the information theoretic model selection, alternative models $f_k(x|\boldsymbol{\theta}_j)$ are compared with the unknown $f(x)$ according to an information discrepancy measure $\mathcal{D}[f(x) : f_k(x|\boldsymbol{\theta}_j)]$. The subscripts emphasize that the models can be in different families and the parameter space can be different. The issue of models being in the same or separate families often arises in statistics. Cox (1961) addressed this issue in testing, which also known as the problem of nested or non-nested hypotheses. Pesaran (1987) operationalized the concept of nested and non-nested hypotheses in terms of the discrimination information in a very precise manner.

Application of (11) in the estimation problem where $f(x)$ is assumed to be in a parametric family $f(x) = f(x|\boldsymbol{\theta})$ is known as the entropy loss (see, e.g., James and Stein, 1961 and Haff, 1980). The objective is to compute an estimate $f(x|\tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}}$ is referred to as the minimum discrimination information or minimum entropy loss estimate.

Akaike (1973) noted that for model selection (13) gives

$$K[f(x) : f_k(x|\boldsymbol{\theta}_j)] = I[f(x)] - I_f[f_k(x|\boldsymbol{\theta}_j)], \tag{32}$$

where $I_f[f_k(x|\boldsymbol{\theta}_j)]$ is defined in (14). Since $I[f(x)]$ is free from $f_k(x|\boldsymbol{\theta}_j)$, the first term in (32) is ignored in the derivation of the AIC criteria for model selection. The AIC type measures are derived by minimizing various estimates of the second term in (32); Akaike information criteria (AIC) uses the sample average of the log-likelihood function for the estimate. Consequently, the AIC type measures do not provide information diagnostics about the model fit. These measures provide criteria for model comparison, irrespective of whether or not the fit of models are satisfactory. Assessing whether the unknown $f(x)$ can be satisfactorily approximated by a parametric model requires estimation of the information discrepancy between the unknown data-generating distribution and the model, $K[f(x) : f_k(x|\boldsymbol{\theta}_j)]$. In general, estimation of the minimum discrimination information function (32) when the data-generating distribution $f(x)$ is unknown constitutes a difficult problem.

Application of (26) alleviates estimation of (32). Let $f(x|\boldsymbol{\theta})$ denote a class of models indexed by $\boldsymbol{\theta}$ for approximating $f \in \Omega_{\boldsymbol{\theta}}$ as defined in (20) and $f^*(x|\boldsymbol{\theta})$ be the ME model in $\Omega_{\boldsymbol{\theta}}$. Using $f^*(x|\boldsymbol{\theta})$ for the parametric model in (32), the second term in (32) becomes $H[f^*(x|\boldsymbol{\theta})]$ and (32) becomes the ID relation (26). This reduces the problem of estimating $K[f(x) : f_k(x|\boldsymbol{\theta}_j)]$ to the problem of estimating the two entropies shown in (26). Then (32) can be estimated by

$$\tilde{K}[f(x) : f_k(x|\boldsymbol{\theta}_j)] = H[f^*(x|\tilde{\boldsymbol{\theta}})] - \tilde{H}[f(x)], \tag{33}$$

where $\tilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ obtained by the moments of the distribution whose entropy is $\tilde{H}[f(x)]$, which is an entropy estimate such as (8). With these moments, $\tilde{K}[f(x) : f_k(x|\boldsymbol{\theta}_j)] \geq 0$ and

provides ID criteria for estimating closeness of the ME model $f^*(x|\tilde{\boldsymbol{\theta}})$ to $f(x)$. Unlike AIC which provides criteria for model comparison purposes only, the ID statistic (33) provides distributional diagnostics for model comparison as well as for the goodness-of-fit. Mazzuchi, et al. (2008) provide applications where $\tilde{H}[f(x)]$ is estimated by the posterior mean of the quantile entropy (8), and give many references. The deviance measure widely used in the exponential family regression is an ID statistic; see, e.g., Hastie (1997) and Spigelhalter et al (2002).

6 Information Measures of Dependence

6.1 Notions of Dependence Information

Two random variables X_1 and X_2 are independent if and only if $f(x_i|x_j) = f(x_i)$ for all $x_i, x_j,\ i \neq j = 1, 2$. Thus for the independent case, for any uncertainty function, $\mathcal{U}[f(x_i|x_j)] = \mathcal{U}[f(x_i)],\ i \neq j = 1, 2$ for all x_j, and for any information discrepancy function $\mathcal{D}[f(x_i|x_j) : f(x_i)] = 0,\ i \neq j = 1, 2$ for all x_j. The condition of for all x_j is sufficient, but not necessary. Since $f(x_i) = E_{x_j}[f(x_i|x_j)]$, for any concave uncertainty function $\mathcal{U}$ we have by Jensen inequality $\mathcal{U}[f(x_i)] \geq E_{x_j}\{\mathcal{U}[f(x_i|x_j)]\}$ where the equality holds if and only the two variables are independent. This leads to the following two measures of information dependence:

$$\mathcal{D}_1(X_1, X_2) = \mathcal{U}[f(x_2)] - E_{x_1}\{\mathcal{U}[f(x_2|x_1)] \geq 0 \tag{34}$$

$$\mathcal{D}_2(X_1, X_2) = \mathcal{U}[f(x_1)] - E_{x_2}\{\mathcal{U}[f(x_1|x_2)] \geq 0, \tag{35}$$

where in each case the equality holds if and only the two variables are independent.

Alternatively, two random variables X_1 and X_2 are independent if and only if $f(x_i, x_j) = f(x_i)f(x_j)$ for all $x_i, x_j,\ i \neq j = 1, 2$. Thus in the independent case, for any information discrepancy function $\mathcal{D}[f(x_i, x_j) : f(x_i)f(x_j)] = 0,\ i \neq j = 1, 2$ for all x_j. This leads to the following measure of information dependence:

$$\mathcal{D}_3(X_1, X_2) = \mathcal{D}[f(x_1, x_2) : f(x_1)f(x_2)] \geq 0, \tag{36}$$

where the equality holds if and only the two variables are independent. Furthermore, if $\mathcal{U}(\cdot)$ is an additive uncertainty measure, then Axiom 4B provides the following measure of information dependence:

$$\mathcal{C}(X_1, X_2) = \mathcal{U}[f(x_1)] + \mathcal{U}[f(x_2)] - \mathcal{U}[f(x_1, x_2)], \tag{37}$$

The measures (36) and (37) are symmetric in X_1 and X_2, but in general (37) can be positive or negative, so it is not an information discrepancy function between $f(x_1, x_2)$ and $f(x_1)f(x_2)$. If $\mathcal{C}(X, Y) \geq 0$, the uncertainty function $\mathcal{U}(\cdot)$ is referred to as *sub-additive*, and then (37) is an information discrepancy function, $\mathcal{C}(X_1, X_2) = \mathcal{D}_4(X_1, X_2)$. In general, the four dependence information measures (34)-(37) are not equal. For example, Kent (1983) defined measures of dependence in terms of Fraser information (14), which can be related to (34), (35), and (37) where $\mathcal{U}(f) = -I_g(f)$. These measures are equal for the bivariate normal case, but not in general; see Kent (1983) and Inaba and Shirahata (1986) for details and examples. Also see, Example 6.3.

The measures (34) and (35) quantify the expected information provided by each variable about the other. The observed information provided by a given $X_j = x_j$ for predicting outcomes of $X_i,\ i \neq j$ is measured by the uncertainty difference

$$\Delta\mathcal{U}[f(x_i|x_j) : f(x_i)] = \mathcal{U}[f(x_i|x_j)] - \mathcal{U}[f(x_i)], \quad i \neq j = 1, 2. \tag{38}$$

For a particular x_j, this measure can be positive, negative, or zero, depending on which one of the two is closer or farther to the uniform distribution. An observation x_j can reduce or increase uncertainty or leave it unchanged; an increase of uncertainty is referred to as "surprise" (Lindley, 1956). In general, $\Delta\mathcal{U}[f(x_i|x_j) : f(x_i)] = 0$ for some x_j, neither implies that the two distributions are identical, nor implies that X_1 and X_2 are independent. The information discrepancy $\mathcal{D}[f(x_i|x_j) : f(x_i)] \geq 0,\ i \neq j = 1, 2$ is also a measure of observed information. For this measure, the equality holds if and only if $f(x_i|x_j) = f(x_i)$ for almost all x_i, but it does not imply the X_1 and X_2 are independent.

6.2 Mutual Information

In terms of Shannon entropy and KL function all four dependence information (34)-(37) are equal and the unique measure is known as the *mutual information* between two random variables:

$$M(X_1, X_2) = \mathcal{D}_1(X_1 : X_2) = \mathcal{D}_2(X_1 : X_2) = \mathcal{D}_3(X_1, X_2) = \mathcal{D}_4(X_1, X_2) \geq 0. \tag{39}$$

It is clear from $\mathcal{D}_3(X_1 : X_2) = K[f(x_1, x_2) : f(x_1)f(x_2)] \geq 0$ that the last equality in (39) holds if and only if X_1 and X_2 are independent, see Cover and Thomas (1991) and Joe (1989). Furthermore, the absolute continuity $F(x_1, x_2) \ll F(x_1)F(x_2)$ is necessary; see Ebrahimi, et al. (2007) for details. The shared information representation $\mathcal{D}_4(X_1, X_2)$ facilitates computation of the mutual information by entropy expressions for many well-known families of distributions.

Mutual information between more than two random variables is defined similarly. For example, the mutual information between all d components of $\boldsymbol{X} = (X_1, \cdots, X_d)$ is given by

$$M(\boldsymbol{X}) = K[f(\boldsymbol{x}) : f_1(x_1) \cdots f_d(x_d)] = \sum_{i=1}^{d} H(X_i) - H(\boldsymbol{X}) \geq 0, \tag{40}$$

where the equality holds if and only if all components are mutually independent. Thus, $M(\boldsymbol{X})$ provides a measure of complexity of a multivariate distribution in terms of dependence between its components, where the simplest multivariate distribution has mutually independent components.

The mutual information measures possess all the properties of (11). For a decomposition similar to (15), let $\mathcal{M}(X_1, X_2|X_3) = E_3\{M(X_1, X_2|X_3)\}$, which measures the conditional dependence and is referred to as the conditional mutual information between X_1 and X_2 given X_3. Successive applications give a chain rule for the mutual information between a random variable Y and a random vector $\boldsymbol{X}$:

$$M(Y, \boldsymbol{X}) = \sum_{i=1}^{d} \mathcal{M}(Y, X_i|X_1, \cdots, X_{i-1}), \tag{41}$$

where $\mathcal{M}(Y, X_i|X_2, \cdots, X_{i-1})$ is the partial mutual information and $\mathcal{M}(Y, X_1|X_1) = M(Y, X_1)$.

Let $(W_1, W_2) = [\phi_1(X_1), \phi_2(X_2)]$, where ϕ_i is one-to-one for all $i = 1, 2$. Then $M(W_1, W_2) = M(X_1, X_2)$. For any copula transformation $C(X_1, X_2)$, the marginal distributions $U_i = C_i(X_i) = F_i(X),\ i = 1, 2$ are uniform over the unit interval, so the marginal entropies are $H(U_i) = 0,\ i = 1, 2$. By the shared information representation, $M(X_1, X_2) = -H(U_1, U_2)$.

Letting $f_1 = f(x_1, x_2)$ and $f_2 = f(x_1)f(x_2)$ in (19), Rényi and Tsallis divergence measures with $\alpha = 1/2$ provide symmetric measures of dependence. Hirschberg, et al. (1991) used $K_{T,1/2}[f(x_1, x_2) : f(x_1)f(x_2)]$ for clustering time series and Granger, et al. (2004) defined a measure in terms of $K_{T,1/2}[f(x_1, x_2) : f(x_1)f(x_2)]$ for detecting non-linear serial dependence of variables. The equivalence of Tsallis divergence $K_{T,1/2}[f(x_1, x_2) : f(x_1)f(x_2)]$ with the Hellinger distance and Bhattacharya distance (19) is appealing. However, unlike the case of Shannon entropy and KL function,

for the generalized entropies and divergence measures, the four dependence information (36) and (37) are not necessarily equal.

Example 6.1 Consider the following distribution

$$f(x_1, x_2, x_3) = \frac{1}{4}, \quad \text{for } (x_1, x_2, x_3) = (0,0,0),\ (0,1,1),\ (1,0,1),\ (1,1,0).$$

This distribution is the classic example for illustrating that pairwise independence does not imply independence. The information method is simple and easily generalizable to higher dimensions. The entropies are $H(X_i) = \log 2,\ H(X_i, X_j) = 2\log 2,\ j \neq i = 1,2,3$, and $H(X_1, X_2, X_3) = 2\log 2$. The shared information formula gives $M(X_i, X_j) = D_4(X_i, X_j) = 0$, so the variables are pairwise independent. The complexity formula (40) gives $M(X_1, X_2, X_3) = \log 2$, so the variables are not mutually independent. The chain rule formula (41) gives $M[X_i, (X_j, X_k)] = \log 2$, so X_i is not independent of (X_j, X_k). Let $Y_{ij} = a_i X_i + a_j X_j, i \neq j = 1,2,3$, then $M(Y_{ij}, X_k) = \log 2,\ k \neq i, j$. Thus, Y_{ij} and X_k are not independent, but they are uncorrelated, due to the fact that the dependence is not linear. By the information method we can easily show this concept more generally with $f(x_1, \cdots, x_n) = \frac{1}{2^{n-1}}$ for $n-1$ points and k-dimensional marginals $f(x_1, \cdots, x_k) = \frac{1}{2^k},\ k \leq n-1$, where $H(X_i) = \log 2,\ H(X_i, X_j) = 2\log 2, \cdots$, and $H(X_1, \cdots, X_n) = (n-1)\log 2$.

Example 6.2 (Location-scale family) Using shared information representation in (39), we find the mutual information between components of a multivariate scale family:

$$M(\boldsymbol{X}; \boldsymbol{\Sigma}) = M(\boldsymbol{X}^*; \boldsymbol{R}) + M(\boldsymbol{X}; \boldsymbol{I}_d) \geq M(\boldsymbol{X}^*; \boldsymbol{R}), \tag{42}$$

where

$$M(\boldsymbol{X}^*; \boldsymbol{R}) = g(\boldsymbol{R}) = -\frac{1}{2}\log|\boldsymbol{R}| = -\frac{1}{2}\sum_{k=1}^{d}\log\lambda_k \geq 0$$

is the multivariate normal mutual information and $\lambda_i,\ i = 1, \cdots, d$ are the eigenvalues of $\boldsymbol{\Sigma}$. Thus, the mutual information (42) decomposes into a measure of linear dependency $M(\boldsymbol{X}^*; \boldsymbol{R})$ and nonlinear dependency $M(\boldsymbol{X}; \boldsymbol{I}_p)$. Among all distributions in the multivariate location-scale family having the same scale matrix $\boldsymbol{\Sigma}$, the multivariate normal model has the minimal dependence structure; it is the least complex distribution. In the family, the multivariate normal with the uncorrelated components attains the global minimum, $M(\boldsymbol{X}^*; \boldsymbol{I}_d) = 0$, mapping the independence. In each family, the distribution with orthogonal components is the least complex. For example, another member of the location-scale family is the multivariate Cauchy distribution with location $\boldsymbol{\mu}$ and scale $\boldsymbol{\Sigma}$. Thus, $M(\boldsymbol{X}; \boldsymbol{I}_d) < M(\boldsymbol{X}; \boldsymbol{\Sigma}) < M(\boldsymbol{X}^*; \boldsymbol{\Sigma})$; see Abe and Rajagopal (2001) for an application. The marginal distributions of subvectors $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ of the multivariate Cauchy are also Cauchy. The mutual information between two disjoint subvectors of the standard multivariate Cauchy variable can be easily computed using representation $M(\boldsymbol{X}_a, \boldsymbol{X}_b; \boldsymbol{I}_d) = \mathcal{D}_4(\boldsymbol{X_a}, \boldsymbol{X_b}; \boldsymbol{I}_d)$. Since the traditional measures (variance and correlation coefficient) are not defined for the Cauchy distribution, the entropy and mutual information are particularly useful.

Rényi entropy of the d-dimensional normal distribution with covariance $\boldsymbol{\Sigma}$ is given by

$$H_{R,\alpha}(\boldsymbol{X}) = \frac{d}{2}\log 2\pi + \frac{d\log\alpha}{2(\alpha-1)} + \frac{1}{2}\log|\boldsymbol{\Sigma}|.$$

Rényi normal entropy is sub-additive. For example, for bivariate normal with correlation ρ,

$$\mathcal{D}_j(X_1, X_2; \alpha) = M(X_1, X_2) = -\frac{1}{2}\log(1-\rho^2) \geq 0, \quad j = 1, 2, 4, \ \forall \alpha > 0$$

is free from α. Rényi dependence information divergence measure for the bivariate normal distribution depends on α and is defined only when $\alpha \leq 1 + \frac{1}{\rho}$. The following example shows the more general case where $\mathcal{D}_j(X_1, X_2; \alpha)$, $j = 1, 2, 3, 4, \alpha \neq 1$ all depend on α, are different, and Rényi entropy is not sub-additive.

Example 6.3 (Gamma-Pareto) Consider the bivariate distribution with density function

$$f(x_1, x_2) = \frac{1}{\Gamma(\beta)} x_1^{\beta} e^{-x_1 - x_1 x_2}, \quad x_1, x_2 \geq 0, \quad \beta > 0. \tag{43}$$

Singpurwalla (2006) referred to (43) as Gamma-Pareto distribution. Darbellay and Vajda (2000) referred to this distribution as Gamma-Exponential and computed Shannon entropy and mutual information for (43). Nadarajah and Zografos (2005) computed its Rényi entropy. The marginal distribution of X_1 is gamma $Ga(\beta, 1)$ with density function

$$f(x_1) = \frac{1}{\Gamma(\beta)} x_1^{\beta-1} e^{-x_1}, \ x_1 > 0, \quad \beta > 0$$

and the marginal distribution of X_2 is Pareto distribution with pdf (47) in Section 8.1. The conditional distribution of $X_2|x_1$ is exponential with rate x_1, and the conditional $f(x_1|x_2) = Ga(\beta + 1, x_2 + 1)$. The KL information can also be easily computed. The information measures are particularly useful since for $\beta \leq 2$, the variance of Pareto and correlation coefficient of (43) are not defined. Using formulas for Rényi entropies of (43), gamma and Pareto distributions, we find $\mathcal{D}_2(X_i, X_j; \alpha)$, $\mathcal{D}_3(X_i, X_j; \alpha)$, and $\mathcal{C}_4(X_i, X_j; \alpha)$. Figure 2 shows the plots of these measures against α. These measures are different for $\alpha \neq 1$ and $\mathcal{C}(X_i, X_j; \alpha) < 0$ for $\alpha \geq 1.39$, thus Rényi entropy of (43) is not subadditive.

7 Bayesian Information Measures

For the Bayesian information measures, X_1 and X_2 of previous sections assume specific interpretations in terms of a vector of observable quantities $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$, an unobservable parameter Θ which can be a scalar or vector, and a quantity to be predicted which we denote by Y_{n+1}. The prior and posterior distributions of Θ are $f(\theta)$ and $f(\theta|\boldsymbol{y})$. The predictive prior and posterior distributions of Y_{n+1} are given by

$$\begin{aligned} f(y_{n+1}) &= \int f(y_{n+1}|\theta) dF(\theta) \\ f(y_{n+1}|\boldsymbol{y}) &= \int f(y_{n+1}|\theta) dF(\theta|\boldsymbol{y}). \end{aligned} \tag{44}$$

Applications of the entropy, KL information, and the mutual information to the prior and posterior distributions of the parameter, and to the predictive distributions produce Bayesian information measures about the parameter Θ and for prediction of Y_{n+1}. Developing prior distributions and model for the likelihood function according to the MEIC and MDIC are well-known.

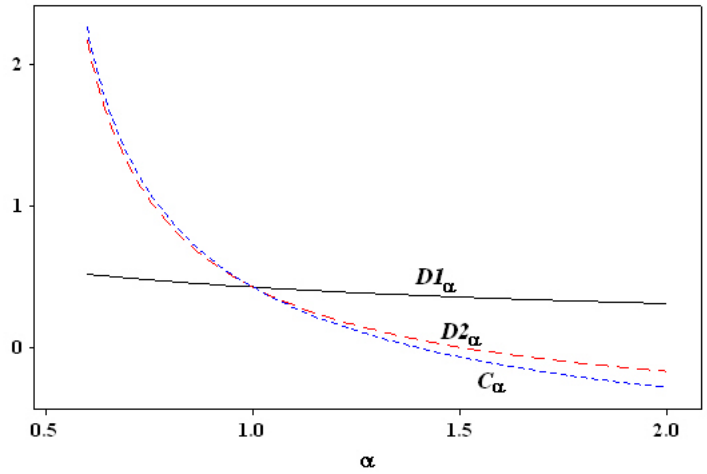

Figure 2: Rényi dependence information measures $\mathcal{D}_2(X_i, X_j; \alpha)$, $\mathcal{D}_3(X_i, X_j; \alpha)$, and $\mathcal{C}(X_i, X_j; \alpha)$ for Gamma-Pareto distribution

The difference (38) between the prior and posterior entropies $\Delta H[f(\theta|\boldsymbol{y}), f(\theta)]$ gives the information provided by the observed sample $\boldsymbol{y}$ about the parameter. Abel and Singpurwalla (1994) utilized the lack of invariance of the observed information for an interesting application. Sample information is also measured in terms of $K[f(\theta|\boldsymbol{y}) : f(\theta)]$, which only detects the change between the prior and the posterior, without indicating which of the two distributions is more informative. The KL information involving predictive distributions have been used by Aitchison (1975, 1990), Johnson and Geisser (1983), Amaral and Dunsmore (1985), Geisser (1993), and Keyes and Levy (1996), among others.

The mutual information $M(\boldsymbol{Y}; \Theta)$ measures the expected sample information about the parameter (Lindley, 1956), and is known as Lindley's measure. Lindley (1961) showed that ignorance between two neighboring values θ and $\Delta\theta$ in the parameter space implies that $M(\boldsymbol{Y}; \Theta) \approx 2(\Delta\theta)^2 \mathcal{F}(\theta)$ where $\mathcal{F}(\theta)$ is Fisher information. A similar result is given by Kullback and Leibler (1951): $K(f_\theta : f_{\theta+\Delta\theta}) \approx 2(\Delta\theta)^2 \mathcal{F}(\theta)$. Polson (1992) developed an approximation of $M(\boldsymbol{Y}; \Theta)$ in the context of nonlinear models. Carota, et al. (1996) in the context of "model elaboration" developed a linearized approximation of $M(\boldsymbol{Y}; \Theta)$ in terms of the Savage density ratio and the score function.

Bernardo (1979a) explicated $M(\boldsymbol{Y}; \Theta)$ as the expected utility when the decision problem is to report a probability distribution $f_p(\cdot)$ from the space of all distributions $\mathcal{P}$ about θ when $\boldsymbol{y}$ is observed. He showed that the logarithmic utility (score) function

$$u[f_p(\cdot), \theta] = A \log f_p(\theta) + B(\theta),$$

where A is a constant and $B(\cdot)$ is an arbitrary function, leads to $M(\boldsymbol{Y}; \Theta) = E_{\theta|\boldsymbol{y}} \{u[f_p(\cdot), \theta]\}$ and thus $f(\theta|\boldsymbol{y})$ as the information optimal distribution for inference about θ. This logarithmic utility function is a member of a large class of utility functions discussed by Good (1971) and others which lead to the posterior distribution given by the Bayes rule as the optimal distribution. Abbas (2004) proposed developing ME and MDI utility functions when partial information about the preference is available.

Zellner (1988) defined an Information Processing Rule (IPR) based on four information measures and derived Bayes rule as the optimal solution. He considered $f(\theta)$ and $f(\boldsymbol{y}|\theta)$ as two *ante data*

distributions for inputs into an information processing which provides two *post data* distributions $f_p(\theta|\boldsymbol{y})$ and $f_p(\boldsymbol{y})$ as outputs. Then Zellner's IPR is defined as:

$$\text{IPR}[f_p(\theta|\boldsymbol{y})] = \overbrace{\left\{I_{f_p}[f_p(\theta|\boldsymbol{y})] + I_{f_p}[f_p(\boldsymbol{y})]\right\}}^{\text{Output information}} - \overbrace{\left\{I_{f_p}[f(\theta)] + I_{f_p}[f(\boldsymbol{y}|\theta)]\right\}}^{\text{Input information}} \tag{45}$$

where $I_{f_p}[\cdot]$ is the information measure defined in (14). Zellner (1988) used calculus of variations and showed that the Bayes rule is the most efficient IPR in the following sense: $\text{IPR}[f_p(\theta|\boldsymbol{y})] = 0$, if and only if $f_p(\theta|\boldsymbol{y}) = f(\theta|\boldsymbol{y})$ is the posterior distribution given by the Bayes rule. Jaynes (discussion of Zellner, 1988) articulated the efficiency of Bayes rule as follows: "An acceptable inference procedure should have the property that it neither ignores any of the input information nor injects any false information; if this requirement already determines Bayes's theorem, the issue seem to be settled." Kullback (discussion of Zellner, 1988) noted that

$$\text{IPR}[f_p(\theta|\boldsymbol{y})] = K[f_p(\theta|\boldsymbol{y}) : f(\theta|\boldsymbol{y})] \geq 0, \tag{46}$$

and the equality gives $f_p(\theta|\boldsymbol{y}) = f(\theta|\boldsymbol{y})$ almost everywhere. (This representation also has been noted by others, see Zellner (Reply 1988) for references). We note that, by the KL representation (46), Zellner's IPR is endowed with the axiomatic foundations of the MDIC. Zellner (Reply 1988) pointed out the possibility of giving different weights to the input components of the IPR and later presented an example of (45) where $f_{\alpha_1}(\theta) \propto f^{\alpha_1}(\theta)$ and $f_{\alpha_2}(\boldsymbol{y}|\theta) \propto f^{\alpha_2}(\boldsymbol{y}|\theta),\ \alpha_j > 0,\ j = 1, 2$; see, Zellner (1997, 2002). Zellner (personal communication, 2010) indicated that the weights α_j's "were introduced by others in the literature to adjust for the quality of the informational inputs ... they were not my invention". The weighted version of IPR also admits a KL representation similar to (46). Zellner (2002) discussed inclusion of side conditions for θ in terms of moments or differential equations which give $f_p(\theta|\boldsymbol{y})$ in the form of the MDI model (21) with reference distribution $f_0(\theta) = f(\theta)$.

Ibrahim, et al. (2003) developed posterior for θ which minimizes

$$K(f) = \alpha K[f(\theta|\boldsymbol{y}) : f_1(\theta|\boldsymbol{y})] + (1-\alpha)K[f(\theta|\boldsymbol{y}) : f_0(\theta|\boldsymbol{y}_0)],\ 0 \leq \alpha \leq 1,$$

where $f_0(\theta|\boldsymbol{y}_0) \propto f_0(\theta)f(\boldsymbol{y}_0|\theta)$ is the posterior which updated the prior $f_0(\theta)$ based on "historic" data $\boldsymbol{y}_0$ before the current data $\boldsymbol{y}$, and $f_1(\theta|\boldsymbol{y}) \propto f_0(\boldsymbol{y}_0|\theta)f(\boldsymbol{y}|\theta)$. They showed that the optimal posterior is given by $f^*(\theta|\boldsymbol{y}) \propto f_0(\theta)f_0^{\alpha}(\boldsymbol{y}_0|\theta)f(\boldsymbol{y}|\theta)$, and $f(\theta) \propto f_0(\theta)f_0^{\alpha}(\boldsymbol{y}_0|\theta)$ is referred to as the power prior. They also included a third component $I_{f_p}[f^{\alpha_3}(\boldsymbol{y}_0|\theta)]$ to the inputs in the weighted version of IPR (45) and showed that $f^*(\theta|\boldsymbol{y})$ is optimal with weights $(\alpha_1, \alpha_2, \alpha_3) = (1, 1, \alpha)$.

Bernardo (1979b) proposed developing reference prior that maximizes $M(\boldsymbol{Y}; \Theta)$. In general, the solution does not have a closed form. Lindley's approximation of $M(\boldsymbol{Y}; \Theta)$ in terms of Fisher information implies that Jeffreys' prior is an approximation to the density maximizing $M(\boldsymbol{Y}; \Theta)$. Bernardo (2004) extended the reference analysis to the symmetric intrinsic information measure $\delta(\Theta, \boldsymbol{y}) = \min\{K[f(\theta, \boldsymbol{y}) : f(\theta)f(\boldsymbol{y})], K[f(\theta)f(\boldsymbol{y}) : f(\theta, \boldsymbol{y})]\}$. Yuan and Clarke (1999) proposed developing models for the likelihood function that maximize $M(\boldsymbol{Y}; \Theta)$ subject to a constraint in terms of the Bayes risk $E_{\theta|\boldsymbol{y}}\{\mathcal{L}(\theta, \boldsymbol{y})\} \leq L_0$, where $\mathcal{L}(\theta, \boldsymbol{y})$ is the loss of using the model $f(\boldsymbol{y}|\theta)$ to learn about the parameter. The optimal solution is the MDI model (21) with $T(\boldsymbol{y}) = \mathcal{L}(\theta, \boldsymbol{y})$.

Lindley's measure has been used by several authors in design problems, both in the context of the normal linear models as well as other models. Stone (1959) was the first to apply $M(\boldsymbol{Y}; \Theta)$ to design of experiments in the context of normal linear models. Following Bernardo (1979a), several authors have presented selection of the optimal design in terms of $M(\boldsymbol{Y}; \Theta)$ as a Bayesian decision

problem; Chaloner and Verdinelli (1995) provide an extensive review and include many references; see also Barlow and Hsiung (1983) and Polson (1993). Briefly, the optimal design problem is cast as a two-stage decision problem where at the first stage a specific design, $X \in \mathcal{X}$, is chosen and data $\boldsymbol{y}$ is observed. Given $\boldsymbol{y}$, the decision at the second stage is which probability density $f_r \in \mathcal{P}$ to report for Θ. Using the logarithmic utility function, maximization of the pre-posterior expected utility $E_{\boldsymbol{y}|X}\left[\sup_{f_r \in \mathcal{P}} E_{\theta|\boldsymbol{y},X}\left\{u[\theta, f_r(\cdot)]\right\}\right]$, where $\boldsymbol{y}$ is the data observed from a specific design (experiment) X, gives the same result as Bernardo's (1979a) formulation (Polson, 1993). Verdinelli, et al. (1993) also proposed optimal design in terms of the predictive mutual information $M(\boldsymbol{Y}; Y_{n+1})$; this measure was used by San Martini and Spezzaferri (1984) for model selection and studied by Amaral and Dunsmore (1985).

An alternative to Lindley's measure for developing priors is the maximal data information prior (MDIP) criterion proposed by Zellner (1977). This measure is the difference between *á priori* average information in the model for data distribution and the information in the prior distribution:

$$\mathcal{Z}(\Theta) = E_\theta[I(Y|\theta)] - I(\Theta),$$

where $I(Y|\theta) = -H(Y|\theta)$. The MDIP maximizes $\mathcal{Z}(\Theta)$ and the solution is in the form of $f(\theta) \propto e^{I(Y|\theta)}$. For a bounded parameter space, the MDIP provides a proper prior distribution, but for an unbounded parameter space the prior can be improper (see Zellner 1997 for details and examples).

Next, we illustrate applications of Bayesian information measures in the regression; details and various other applications of information measures to regression are given in Soofi (1990, 1997).

Example 7.1 Consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is an $n \times 1$ vector of observed values, $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{\theta}$ is the $p \times 1$ parameter vector, and $\boldsymbol{\epsilon}$ is $n \times 1$ vector of random errors. The ME model for likelihood function is obtained using a variation constraint on the error term. Under the square error constraint $E(\epsilon_i^2) \leq \sigma^2$, the ME model is independent multivariate normal $f^*_{\boldsymbol{\epsilon}|\sigma^2} = N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. The prior is obtained by the ME model in the location-scale family $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma_0^2 C_0$, which is the multivariate normal $f^*(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_0, \sigma_0^2 \boldsymbol{C}_0)$.

The optimal design problem involves the choice of design matrix $\boldsymbol{X}$. Lindley's information is

$$M(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{X}) = \Delta H[f(\boldsymbol{\theta}|\boldsymbol{y}), f(\boldsymbol{\theta})] = \frac{1}{2} \log \left| \eta^{-1} \boldsymbol{C}_0 \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p \right|,$$

where $\eta = \sigma^2/\sigma_0^2$. Thus, the information optimal design is $X^* = \arg\max_{X \in \mathcal{X}} M(\boldsymbol{Y}; \boldsymbol{\Theta}|\boldsymbol{X})$. If $\boldsymbol{C}_0 = \boldsymbol{I}_p$, the optimal design X^* is orthogonal, which is $D-$optimal in the classical design literature. The classical $D-$optimality is also obtained when the prior is weak, $\sigma_0^2 \to \infty$. In the limit, the prior is uniform and $M(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{X})$ is not defined. In this case, $f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) = N\left(\boldsymbol{b}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2\right)$, where $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the least squares estimate and the posterior information about the parameter is given by the negative multivariate normal entropy (29), so $I(\boldsymbol{\Theta}|\boldsymbol{y}) \propto \log |\boldsymbol{X}'\boldsymbol{X}|^{-1}$.

Measures of information loss due to collinearity compare the posterior distribution based on the actual regression matrix $\boldsymbol{X}$ and the posterior distribution for the optimal design $\boldsymbol{X}^*$ (Soofi, 1990). In terms of choice of prior, the maximum sample information is attained when $\boldsymbol{C}_0 \propto (\boldsymbol{X}'\boldsymbol{X})^{-1}$, which is the case of some noninformative priors such as the reference prior, and Zellner's g prior. With such priors, the posterior covariance structure remains the same and leads to unreliable inference when the collinearity is severe. Informative priors compensate for the collinearity effects.

Influence diagnostics are obtained by information discrepancy between inferential distributions based on all observations and based on deletion of some. Let $\boldsymbol{X}_{-i}$ and $\boldsymbol{y}_{-i}$ denote the data with the ith observation deleted. For example, under noninformative prior, the change in the amount

of uncertainty in predicting a value of $\boldsymbol{\theta}$ due to the presence and absence of the ith observation is given by the posterior entropy difference

$$\Delta H[f(\theta|\boldsymbol{y}), f(\theta|\boldsymbol{y}_{-i})] = -\log\left(|\boldsymbol{X}'\boldsymbol{X}|^{-1}|\boldsymbol{X}'_{-i}X_{-i}|\right) = -\frac{1}{2}\log(1-h_{ii}) \geq 0,$$

where h_{ii} is the ith diagonal element of the projection matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$. For influence information measures about the parameter using (11), see Soofi (1997).

Johnson and Geisser (1983) developed influence diagnostics for predictive distribution of n new observations $\boldsymbol{y}_{new}$ to be taken at the regression matrix $\boldsymbol{X}$. For example, the predictive influence of an observation is measured by the following discrimination information functions of the predictive densities: $K[f(\boldsymbol{y}_{new}|\boldsymbol{y}_{-i}) : f(\boldsymbol{y}_{new}|\boldsymbol{y})]$ or $K[f(\boldsymbol{y}_{new}|\boldsymbol{y}) : f(\boldsymbol{y}_{new}|\boldsymbol{y}_{-i})]$. Carlin and Polson (1991) developed influence diagnostics using Lindley's measure.

8 Four Application Areas

8.1 Duration Analysis

Study of duration is a subject of interest common to reliability, survival analysis, actuary, economics, business, and many other fields. Information theoretic methods for duration analysis are mainly developed in the context of reliability; see Ebrahimi and Soofi (2004), Singpurwalla (2006), and references therein. This section focuses on the dynamic information measures, where the current "age" becomes a parameter of the model. When the subject of duration study is other than lifetime (e.g., search time, unemployment period) the present time point plays the role of "age".

Let X be a non-negative random variable representing the lifetime of an item and $t \geq 0$ denote its current age. At age t, the pdf of the residual lifetime of the item is $f(x;t) = \dfrac{f(x)}{\bar{F}(t)}$, $x > t$. The residual entropy is given by $H(X;t) = H[f(x;t)]$. Let f^* denote the uniform pdf on $[0,\beta]$. Then conditional (truncated) distribution of X, given $X > t$ is also uniform over $(t,\beta]$. Thus, $H[f(x;t)]$ is the measure of uncertainty of the residual distribution with pdf $f(x;t)$. Considering $\mathcal{S}_t = \{x : x > t\}$ as an index set, $H(X;t)$ is a dynamic uncertainty measure ranging over $\mathcal{S}_t$.

Dynamic discrimination information $K(f_1 : f_2;t) = K[f_1(x;t) : f_2(x;t)]$ and dynamic mutual information $M(X_1, X_2; t_1, t_2)$ are defined similarly. These measures can be used for discrimination and study of dependence between the remaining lifetimes of the components of a system when they have already survived to times t_1, t_2. They are local measures that can be used, for example, to study early/late time dependence. Dynamic information measures for the past lifetime with support $\mathcal{S}_{[t]} = \{x : x \leq t\}$ are defined similarly. The partitioning transformation (1) is applicable to the $\mathcal{S} = \mathcal{S}_t \cup \mathcal{S}_{[t]}$. Consideration of the age has led to some important insights about lifetime models; see Ebrahimi (1996), Ebrahimi and Kirmani (1996), Di Crescenzo and Longobardi (2002, 2004), Ebrahimi, et al. (2007), and references therein.

The dynamic extensions of the MDIC and MEIC are in terms of dynamic information measures $K(f : f_0;t)$ and $H(f;t)$. It is natural to think of the dynamic constraints on the residual moments and hazard rate. However, the mean residual and hazard rate equality constraints uniquely determine the distribution. Asadi, et al. (2004, 2005) considered $\Omega_Q = \{f\}$ where f is subject to the inequality or differential inequality constraints for the hazard rate or inequality constraints on the mean residual. The dynamic information optimal models are defined as follows.

Definition 3 *The MDDI model in Ω_Q reference to f_0 is $f^* = \arg\min_{f\in\Omega_Q} K(f : f_0;t)$, $\forall t \geq 0$.*

Definition 4 *The MDE model in* Ω_Q *is* $f^* = \arg\max_{f\in\Omega_Q} H(f;t), \quad \forall t \geq 0.$

That is, the MDDI (MDE) model $f^* \in \Omega_Q$ is the one that its residual pdf $f^*(x;t)$ retains the MDI (ME) property among all the residual distributions $f(x;t)$ induced by all $f \in \Omega_Q$ for all $t \geq 0$.

The MDDI and MDE models are found by orderings of the dynamic information functions. Thus, the optimal models are boundary solutions. Many distributions are characterized as MDDI and MDE models by applying the facts that hazard ordering implies entropy ordering if f is monotone and likelihood ratio ordering implies dynamic discrimination ordering if $\dfrac{f(x)}{f_0(x)}$ is monotone.

Next example illustrates the ME, MDE, and MDDI characterizations of a Pareto distribution.

Example 8.1 Consider the Pareto distribution pdf

$$f(x) = \beta(1+x)^{-(\beta+1)}, \quad x \geq 0, \ \beta > 0. \tag{47}$$

This pdf can be written in the form of (23) with $T_1(x) = \log(1+x)$, so it is the ME model in $\Omega_\theta = \{f(x|\theta) : E_f[\log(1+x)] = \theta\}$. The pdf (47) is the MDE model in the class of distributions with hazard rate constraints $\Omega_1 = \left\{ f : \dfrac{\lambda_f'(t)}{\lambda_f(t)} \geq -\dfrac{1}{\beta}\lambda_f(t), \ \lambda_f(0) = \beta \right\}$. It is also the MDDI model relative to the exponential reference $f_0(x) = \lambda e^{-\lambda x}$ in $\Omega_2 = \left\{ f : \dfrac{\lambda_f'(t)}{\lambda_f(t)} \leq -\dfrac{\lambda_f(t)}{\beta}, \ \lambda_f(0) = \beta, \ 1 \leq \beta \leq \lambda - 1 \right\}$. More details and examples are given in Asadi, et al. (2004, 2005).

8.2 Order Statistics

Order statistics are used in a wide range of problems in statistics, reliability analysis, quality control, economics, engineering, among others; see (Arnold, 1992). Following, Wong and Chen (1990) and Park (1995), Ebrahimi, et al. (2004) explored information properties of order statistics. Several authors have followed suit. In this section we list a few information properties of order statistics.

Let $Y_1 \leq \cdots \leq Y_n$ denote the order statistics of random variables $X_1, \cdots, X_n$ which have identical distribution $f(x|\theta)$ and given θ, are independent. By the probability integral transformation, $U_i = F(X_i|\theta), \ i = 1, \cdots n$ are samples from the uniform distribution over $[0,1]$. Then $W_i = F(Y_i|\theta), \ i = 1, \cdots n$ are order statistics of the uniform sample and have beta distributions $g_i = Beta(i, n-i+1)$.

(a) The transformation is one-to-one, thus by (7) the entropy of order statistics is given by

$$H(Y_i) = H_n(W_i) - E_{g_i}\left[\log f\left(F^{-1}(W_i)|\theta\right)\right],$$

where $H_n(W_i)$ denotes the entropy of the beta distribution.

(b) The discrimination information between the distribution of order statistics $f(y_i|\theta)$ and the parent distribution $f(x|\theta)$ is distribution free, and is given by $K_n[f(y_i|\theta) : f(x|\theta)] = -H_n(W_i)$. The median has the closest distribution to the parent distribution.

(c) For any pair of order statistics $K_n[f(y_i|\theta) : f(y_j|\theta)]$ and $M_n(Y_i; Y_j|\theta)$ are also distribution free. $M_n(Y_r; Y_{r+1}|\theta)$ measures the Markovian dependence between order statistics of the independent sample conditional on θ. $M_n(Y_r; Y_{r+1}|\theta)$ is increasing in n, and for a given n, the information is symmetric in r and $n-r$, and attains its maximum at the median.

8.3 Data Disclosure

The data disclosure problem is one aspect of the general problem of preserving confidentiality in data analysis. It comes about because in certain societies, notably the U.S., data that is gathered using taxpayer resources has to be made available to the public, but under the caveat that the released data does not betray public trust by compromising confidentiality. As a consequence, government agencies strike a balance by "masking" the data prior to its release, but in a manner that endeavors to preserve the essential information that the data contains.

Karr, et al. (2006) compared data utility in terms of the KL divergence between the empirical distributions of the original and the released data. Keller-McNulty, et al. (2005) proposed a decision-theoretic framework by looking at the problem from the perspective of a data-collection agency. The data-collection agency's utility function consists of balancing the disclosure risk and data utility. The intruder's utility represents the disclosure risk and this is reflected by a function of the entropy of the distribution of the released data. In a similar manner, the legitimate user's utility represents the data utility which is reflected by a function of the entropy difference between the distributions of the released and the original data. Polettini (2003) describes a maximum entropy-based approach for finding a distribution from which the data to be released can be simulated. These developments can be integrated in a comprehensive information theoretic framework for data disclosure. Challenging problems include (a) a general algorithm for computing the parameters of (23) based on multivariate moments of the actual data; (b) application of (33) for testing the compatibility of the ME distribution with a nonparametric multivariate distribution; (c) methods for simulating new data from the ME model (23) for disclosure; and (d) methods to assess preservation of the dependence structure of the actual data in the released data.

8.4 Importance of Predictor Variables

Relative importance measures refer to quantities that compare the contributions of individual explanatory variables to a response variable. A study of scientific literature by Kruskal and Majors (1989) revealed widespread interest in assigning relative importance to explanatory variables in most fields. However the authors concluded that:

> "We were depressed by the frequency of use of statistical significance as a measure of relative importance. Even though we had half expected that misuse, it was sad to see significance testing so often and inappropriately employed" (Kruskal and Majors 1989).

The interpretation of statistical significance in terms of importance which is also seen in some statistics textbooks is unfounded. Statistical significance maps one's strength of confidence about an inference, whereas relative importance measures are magnitudes of some functions of the parameters. The may be known (e.g., population data, simulation study), so inference is irrelevant, but they are usually unknown and are subject to inference. Soofi, et al. (2000) summarized the relative importance literature and concluded that additivity of the predictors' joint importance in terms of their individual shares, and order-independence are two desirable properties.

Theil and Chung (1988) introduced the normal mutual information measure in the relative importance literature, and justified it in terms of the additive decomposition (41). Soofi (1992) defined importance of predictors for logit model in terms of the maximum entropy difference (27). Retzer, et al. (2009) conceptualized importance in terms of the information provided by a predictor for reducing uncertainty about predicting the outcomes. The mutual information is used for stochastic predictors and (27) is used for nonstochastic predictors. Distributions with densities in the exponential family having finite entropies are ME models (Ebrahimi, et al., 2008). For the

exponential family regression, the deviance gives an estimate of the entropy reduction and provides a measure of the information importance of predictors. Retzer, et al. (2009) provided algorithms for Bayesian inference about the information measures of normal regression, contingency tables, and general logit analysis, and many references. The stage is now set for measuring information importance in terms of the Bayesian predictive distribution (44) and developing algorithms for information importance of predictors for other exponential family models, time series, and so on.

8.5 Additional References

This paper provided an overview of several information measures and applications of Shannon entropy and KL information to an assortment of probability and statistics problems in a unified manner. The paper is not exhaustive due to the breadth of the topic. Other papers and books include unification of various models and methods in actuarial science (Brockett, 1991), marketing (Brockett, et al., 1995), regression analysis (Soofi, 1997), process control (Alwan, et al., 1998), model selection (Burnham and Anderson, 1998), and time series (Pourahmadi and Soofi, 2000). Books and special issues of econometric journals are devoted to information theoretic methods include Theil (1967), Theil and Fiebig (1984), Golan, et al. (1996), Fomby and Hill (1997), *Journal of Econometrics* (2002, 2007), *Econometric Reviews* (2008), and Golan (2006). Bozdogan (1994), Cover and Thomas (1991) and Kapur (1989) present many engineering applications and provide details about information measures. A high potential area of application is system design where the problem is formulated in terms of utility functions (Singpurwalla, 1992), so information measures can be used as the utility functions. Other high potential areas of applications include genetics and computational biology where information measures are already in use (Ewens and Grants, 2005).

Acknowledgments

We thank the editor, Eugene Seneta, a reviewer, and our colleague Paul Nystrom for their comments and suggestions which led us to improve the exposition. Ehsan Soofi's research was partially supported by a Lubar School's Business Advisory Council Summer Research Fellowship.

References

Abbas, A.E. (2006). Maximum entropy utility. *Oper. Res.*, **54**, 277-290.

Abe, S. & Rajagopal, A. K. (2001). Information theoretic approach to statistical properties of multivariate CauchyLorentz distributions. *J. Phys. A: Math. and General*, **34**, 8727-8731.

Abel, P.S. & Singpurwalla, N.D. (1994). To survive or to fail: that is the question. *Amer. Statist.*, **48**, 18-21.

Aitchison, J . (1975). Goodness of prediction fit. *Biometrika*, **62**, 547-554.

Aitchison,J .(1990). On Coherence in parametric density estimation. *Biometrika*, **77**, 905-908.

Alwan, L. C., Ebrahimi, N., & Soofi, E.S. (1998). Information-theoretic framework for statistical process control. *European Journal of Operational Research*, **111**, 526-541.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, Eds. B.N. Petrov & F. Csaki, pp. 267-281, Budapest: Akademiai Kiado.

Amaral-Turkman, M.A. & Dunsmore, I. (1985). Measures of information in the predictive distribution. *Bayesian Statistics,* **2**, Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith, pp. 603-612, Amsterdam: Elsevier.

Arnold, B.C., Balakrishnan, N. & Nagaraja, H.N. (1992). *A first course in order statistics*, New York: Wiley.

Asadi, M., Ebrahimi, N., Hamedani, G.G., & Soofi, E.S. (2004). Maximum dynamic entropy models. *J. Appl. Probab.*,**41**, 379-390.

Asadi, M., Ebrahimi, N., Hamedani, G.G., & Soofi, E.S. (2005). Dynamic minimum discrimination information models. *J. Appl. Probab.*, **42**, 643-660.

Barlow, R.E. & Hsiung, J.H. (1983). Expected information from a life test experiment. *Statistician*, **48**, 18-21.

Barron, A. R. (1986). Entropy and the Central Limit Theorem. *Ann. Probab.*, 14, 936-342.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.*, **7**, 686-690.

Bernardo, J. M. (1979b). Reference posterior distribution for Bayesian inference. *J. R. Statist. Soc. Ser. B*, **41**, 605-647 (with discussion).

Bernardo, J. M. (2005). Reference analysis. in *Handbook of Statistics* **25**. Eds. D. K. Dey & C. R. Rao, pp. 17-90, Amsterdam: Elsevier.

Bernardo, J. M. & Rueda, R. (2002). Bayesian hypothesis testing: a reference approach. *International Statistics Review*, **70**, 351–372.

Bozdogan, H. (1994). *Engineering & Scientific Applications of Informational Modeling, Vol. 3, Proceedings of First US/Japan Conference on The Frontiers of Statistical Modeling: An Informational Approach*, Netherlands: Kluwer.

Brockett, P. L. (1991). Information theoretic approach to actuarial science: A unification and extension of relevant theory and applications. *Trans. Soci. Actuaries*, **43**, 73-135.

Brockett, P. L., Charnes, A., Cooper, W. W., Learner, D., & Phillips F. Y. (1995). Information theory as a unifying statistical approach for use in marketing research. *European Journal of Operational Research*, **84**, 310-329.

Burbea, J. & Rao, C.R. (1982). On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Info. Theory*, **28**. 489-495.

Burnham, K. P. & Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*, New York: Springer-Verlag.

Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statist. Sci.*, **10**, 273-304.

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

Carlin, B.P. & Polson, N.G. (1991). An expected utility approach to influence diagnostics. *J. Amer. Statist. Assoc.*, **87**, 1013-1021.

Carota, C., Parmigiani, G. & Polson, N.G. (1996). Diagnostic measures for model criticism. *J. Amer. Statist. Assoc.*, **91**, 753-762.

Cox, D.R. (1961). Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium*, 1, pp.105-123. Berkeley: UC Press.

Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. R. Statist. Soc. Ser. B*, **46**, 440-464.

Csiszar, I. (1975). I-divergence geometry of probability distributions. *Ann. Probab.*, **3**, 146-159.

Csiszar, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference in linear inverse problems. *Ann. Statist.*, **19**, 2032-2066.

Darbellay, G.A. & Vajda, I. (2000). Entropy expressions for multivariate continuous distributions. *IEEE Trans. Info. Theory*, **46**, 709-712.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *Ann. Math. Statist.*, **33** 404-419.

Di Crescenzo, A. & Longobardi, M. (2002). Entropy-based measure of uncertainty in past lifetime distributions. *J. Appl. Probab.*, **39**, 434-440.

Di Crescenzo, A. & Longobardi, M. (2004). A Measure of discrimination between past life-time distributions. *Statist. Probab. Lett.*, **67**, 173-182.

Ebrahimi, N. (1996). How to measure uncertainty in the residual lifetime distributions. *Sankhya A*, **58**, 48-57.

Ebrahimi, N. & Kirmani, S.N.U.A. (1996). A Characterization of the proportional hazards model through a measure of discrimination between two residual life distributions. *Biometrika*, 83, 233-235.

Ebrahimi, N. & Soofi, E. S. (2004). Information measures for reliability. *Mathematical Reliability: An Expository Perspective*, Eds. R. Soyer ,T.A. Mazzuchi, & N. D. Singpurwalla, pp. 127-159. Netherlands: Kluwer.

Ebrahimi, N., Kirmani, S.N.U.A., & Soofi, E.S. (2007). Dynamic multivariate information. *J. Multivariate Anal.*, **98**, 328-349.

Ebrahimi, N., Maasoumi, E., & Soofi, E.S. (1999). Ordering univariate distributions by entropy and variance. *J. Econometrics*, **90**, 317-336.

Ebrahimi, N., Soofi, E.S., & Soyer, R. (2008). Multivariate maximum entropy identification, transformation, and dependence. *J. Multivariate Anal.*, **99**, 1217-1231.

Ebrahimi, N., Soofi, E.S., & Zahedi, H. (2004). Information properties of order statistics and spacings. *IEEE Trans. Inform. Theory*, **50**, 177-183.

Esteban M.D., & Morales, D. (1995). A summary on entropy statistics. *Kybernetica*. **31**, 337-346.

Ewens, W.J. & Grant, G.R. (2005). *Statistical Methods in Bioinformatics: An Introduction, Second ed.*, New York: Springer.

Fraser, D.A.S. (1965). On information in statistics. *Ann. Math. Statist.*, **36**, 890-896.

Fomby, T. B. & Hill, R. C. (1997). *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, **12**, New York: JAI Press.

Gabriele K-I. (1999). A note on conditional logics and entropy. *International J. Approximate Reasoning*, **19**, 231-246.

Geisser, S. (1993). *Predictive Inference: An Introduction*, New York: Chapman-Hall.

Goel, P.K. & DeGroot, M.H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.*, **76**, 140-147.

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

Golan, A. (2006). Information and entropy econometrics–A review and synthesis. *Foundations and Trends in Econometrics*, **2**, 1-145.

Golan, A., Judge, G. G., & Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, New York: Wiley.

Good, I. J. (1971). Discussion of "measuring information and uncertainty". By R.J. Buehler, in *Foundations of Statistical Inference*, Eds. V.P. Godambe, & D.A. Sprott, pp. 337-339. Toronto: Holt, Rinehart & Winston.

Granger, C. W., Maasoumi, E, & Racine J. (2004). A dependence metric for possibly nonlinear processes. *J. Time Ser. Anal.*, **25**, 649-669.

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.*, **8**, 586-597.

Hastie, T. (1987). A Closer Look at the Deviance, *Amer. Statist.*, **41**, 16-20.

Hirschberg, J., Maasoumi, E., & Slottje, D.J. (1991). Cluster analysis and the quality of life across countries. *J. Econometrics*, **50**, 131-150.

Ibrahim, J.G., Chen, M-H., & Sinha, D. (2003). On optimality of the power prior. *J. Amer. Statist. Assoc.*, **98**, 204-213.

Imbens. G., Johnson, P., & Spady, R. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, 66, 333-357.

Inaba, T. & Shirahata, S. (1986). Measures of dependence in normal models and exponential models by information gain. *Biometrika*, **73**, 345-352.

James, W. & C. Stein (1961). Estimation with quadratic loss function. *Proceedings of the Fourth Berkeley Symposium*, **1**, 361-375, Berkeley: UC Press.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620-630.

Jaynes, E. T. (1968). On the rationale of maximum-entropy methods. *Proc. IEEE*, **70**, 939-952.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proc. R. Soc. A*, **186**, 453-461.

Johnson, W. & S. Geisser (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Assoc.*, **78**, 137-144.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer.Statist. Assoc.*, **84**, 157-164.

Jizba, P., & Arimitsu, T. (2004). The world according to Rényi:thermodynamics of multifractal systems. *Ann. Phys.*, **312**, 17-59.

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

Kapur, J. N. (1994). *Measures of Information and Their Applications*, New Dehli: Wiley.

Karbelkar, S.N. (1986). On the axiomatic approach to the maximum entropy principle of inference. *Pramana-J. Phys.*, **26**, 301-310.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. & Sanil, A. P. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *Amer. Statist.*, **60**, 224-232.

Keller-McNulty, S., Nakhleh. C.W. & Singpurwalla, N. D. (2005). A Paradigm for Masking (Camouflaging) Information. *International Statistical Review*, **73**, 331-349.

Kent, J.T. (1982). Robust properties of likelihood ratio test. *Biometrika*, **69**, 19-27.

Kent, J.T. (1983). Information gain and a general measure of correlation. *Biometrika*, **70**, 163-173.

Keyes, T.K. & Levy, M.S. (1996). Goodness of prediction fit for multivariate linear models. *J. Amer. Statist. Assoc.*, **91**, 191-197.

Khinchin, A.I. (1957). *Mathematical Foundations of Information Theory*, New York: Dover.

Kruskal, W. & Majors, R. (1989). Concepts of relative importance in scientific literature. *Amer. Statist.*, **43**, 2-6.

Kullback, S. (1954). Certain inequalities in information theory and the Cramer-Rao inequality. *Ann. Math. Statist.* **25**, 745-751.

Kullback, S. (1959). *Information Theory and Statistics*, N.Y.: Wiley (reprinted in 1968 by Dover).

Kullback, S. (1987). The Kullback-Leibler distance. *Amer. Statist.*, **41**, 340.

Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79-86.

Lindley, D.V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.*, **27**, 986-1005.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp.*, **1**, pp. 436-468, Berkeley: UC Press.

Maasoumi, E. (1993). Compendium to information theory in economics and econometrics, *Econometric Reviews*, **12**, 137-181.

McCulloch, R. E. (1989). Local model influence. *J. Amer. Statist. Assoc.*, **84**, 473-478.

Mazzuchi, T. A., Soofi, E.S., & Soyer, R. (2008). Bayes estimate and inference for entropy and information index of fit. *Econometric Reviews*, **27**, 428-456.

Nadarajah, S. & Zografos, K. (2003). Formulas for Rényi information and related measures for univariate distributions. *Inform. Sci.*, **155**, 119138.

Nadarajah, S. & Zografos, K. (2005). Expressions for Rényi and Shannon entropies for bivariate distributions. *Inform. Sci.*, **170**, 173189.

Park, S. (1995). The entropy of consecutive order statistics. *IEEE Trans. Inform. Theory*, **41**, 2003-2007.

Pesaran, M. H. (1987) "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power", *Econometric Theory*, 3, 69-97.

Polson, N. G. (1992). On the expected amount of information from a nonlinear model. *J. R. Statist. Soc. Ser. B*, **54**, 889-895.

Polson, N. G. (1993). A Bayesian perspective on the design of accelerated life tests. *Advances in Reliability*, Ed. A. P. Basu, pp. 321-330. North Holland.

Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statist. Comput.*, **13**, 307-320.

Pourahmadi, M. & Soofi, E.S. (2000). Predictive variance and information worth of observations in time series. *J. Time Ser. Anal.*, **21**, 413-434.

Retzer, J.J., Soofi, E.S., & Soyer R. (2009). Information importance of predictors: Concepts, measures, Bayesian inference, and applications. *Camput. Statist. Data Anal.*, **53**, 2363-2377.

Rényi, A. (1961). On measures of entropy and information. *Proc. Fourth Berkeley Symp.*, **1**, pp.547-561. Berkeley: UC Press.

Ryu, H. K. (1993). Maximum entropy estimation of density and regression functions. *J. Econometrics*, **56**, 397-440.

San Martini, A. & Spezzaferri, F. (1984). A predictive model selection criteria. *J. R. Statist. Soc. Ser. B*, **46**, 296-303.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 379-423.

Shore, J. E., & Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory*, **26**, 26-37.

Singpurwalla, N.D. (1992). A Bayesian perspective on Taguchi's approach to quality engineering and tolerance design. *Institute of Industrial Engineering Transactions*, **24**, 18-31 (with discussion).

Singpurwalla, N.D. (2006). *Reliability and Risk: A Bayesian Perspective*, New York: Wiley.

Soofi, E.S. (1990). Effects of collinearity on information about regression coefficients. *J. Econometrics*, **43**, 255-274.

Soofi, E.S. (1992). A generalizable formulation of conditional logit with diagnostics. *J. Amer. Statist. Assoc.*, **87**, 812-816.

Soofi, E.S. (1994). Capturing the intangible concept of information. *J. Amer. Statist. Assoc.*, **89**, 1243-1254.

Soofi, E.S. (1997). Information theoretic regression methods. *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, **12**, Eds. T. B. Fomby & R. C. Hill, pp .25-83, New York: JAI Press.

Soofi, E. S. & Retzer, J.J. (2002). Information indices: unification and applications. *J. Econometrics*, **107**, 17-40.

Soofi, E.S., Ebrahimi, N., & Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *J. Amer. Statist. Assoc.*, **90**, 57-668.

Soofi, E. S., Retzer, J.J., & Yasai-Ardekani, M. (2000). A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences*, **31**, 595-625.

Spigelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. Ser. B*, **64**, 583-616 (with discussion).

Stone, M. (1959). Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Statist.*, **29**, 55-70.

Teitler, S., Rajagopal, A.K., & Ngai, K.L. (1986). Maximum Entropy and reliability distributions. *IEEE Trans. Reliab.* **35**, 391-395.

Theil, H. (1967). *Economics and Information Theory.* Amsterdam: North-Holland.

Theil, H. & Chung C. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. *Amer. Statist.*, **42**, 249-252.

Theil, H. & Fiebig, D. Z. (1984). *Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions*, Cambridge, MA: Ballinger.

Tsallis, C. (1988). Possible generalization of Bolzmann-Gibbs statistics. *J. Statist. Phys.*, **52**, 479-487.

Tsallis, C. (1998). Generalized entropy-based criterion for consistent testing. *Phys. Rev. E*, **58**, 1442-1445.

Verdinelli, I., Polson, N.G. & Singpurwalla, N.D. (1993). Shannon information and Bayesian design for prediction in accelerated life-testing. *Reliability and Decision Making*, Eds. R.E. Barlow, C.A. Clarotti, & F. Spizzichino, pp. 247-256. London: Chapman Hall.

Vasicek, O. (1976). A test for normality based on sample entropy. *J. R. Statist. Soc. Ser. B*, **38**, 54-59.

Wong, K.M. & Chen, S. (1990). The entropy of ordered sequences and order statistics. *IEEE Trans. Inform. Theory*, **36**, 276-284.

Yuan, A. & Clarke, B. (1999). An information criterion for likelihood selection. *IEEE Trans. Inform. Theory*, **45**, 562-571.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley (reprinted in 1996 by Wiley).

Zellner, A. (1977). Maximal data information prior distributions. in *New Developments in the Applications of Bayesian Methods*. Eds. A. Aykac & C. Brumat, pp. 211-232, Amsterdam: North Holland.

Zellner, A. (1988). Optimal information processing and Bayes' Theorem. *Amer. Statist.*, **42**, 278-284 (with discussion).

Zellner, A. (1996). Bayesian method of moments/instrumental variable (BMOM/IV) analysis of mean and regression models. In *Prediction and Modelling Honoring Seymour Geisser*, Eds. J.C. Lee, A. Zellner & W.O. Johnson, pp. 61-74. Netherlands: Springer-Verlag.

Zellner, A. (1997). *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*. Cheltenham UK: Edward Elgar.

Zellner, A. (2002). Information processing and Bayesian analysis. *J. Econometrics*, **107**, 41-50.

Zografos, K. & Nadarajah, S. (2005). Expressions for Rényi and Shannon entropies for multivariate distributions. *Statist. Probab. Lett.*, **71**, 71-84.