

BAYES ESTIMATE AND INFERENCE FOR ENTROPY AND INFORMATION INDEX OF FIT

Thomas A. Mazzuchi,¹ Ehsan S. Soofi,² and Refik Soyer³

¹*School of Engineering and Applied Science, George Washington University, Washington, District of Columbia, USA*

²*Sheldon B. Lubar School of Business and Center for Research on International Economics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA*

³*Department of Decision Sciences, George Washington University, Washington, District of Columbia, USA*

□ *This article defines a quantized entropy and develops Bayes estimates and inference for the entropy and a Kullback–Leibler information index of the model fit. We use a Dirichlet process prior for the unknown data-generating distribution with a maximum entropy candidate model as the expected distribution. This formulation produces prior and posterior distributions for the quantized entropy, the information index of fit, the moments, and the model parameters. The posterior mean of the quantized entropy provides a Bayes estimate of entropy under quadratic loss. The consistency of the Bayes estimates and the information index are shown. The implementation and the performances of the Bayes estimates are illustrated using data simulated from exponential, gamma, and lognormal distributions.*

Keywords Dirichlet process; Kullback–Leibler; Model selection; Nonparametric Bayes.

JEL Classification C11; C13; C14; C16; C52.

1. INTRODUCTION

Entropy and Kullback–Leibler information have been instrumental in the development of indices of fit of parametric models to the data. Frequentist inference procedures about entropy-based fit indices are abundant. Although use of a parametric model that fits the data is of paramount importance for Bayesian analysis, entropy-based fit indices and Bayesian inference about them have not received much attention.

Received June 23, 2006; Accepted April 16, 2007

Address correspondence to Ehsan S. Soofi, Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P.O. Box 742, Milwaukee, WI 53201, USA; E-mail: esoofo@uwm.edu

In Bayesian statistics, Kullback–Leibler information and entropy are mainly used as descriptive measures for comparison of parametric models and many other purposes and as criteria for derivation of models and priors; see, e.g., Bernardo (1979), Clarke (1996), Clarke and Gustafson (1998), Yuan and Clarke (1999), Carota et al. (1996), and Zellner (1991, 1996).

Given data x_1, \dots, x_n from a distribution F , we wish to assess whether the unknown $F(x)$ can be satisfactorily approximated by a parametric model $F^*(x|\theta)$. We measure the loss of approximating $F(x)$ by a parametric model $F^*(x|\theta)$ using the Kullback–Leibler discrimination information between the two distributions,

$$K(F : F^* | \theta) = \int \log \frac{f(x)}{f^*(x|\theta)} dF(x), \quad (1)$$

where f and f^* are the respective probability density (mass) functions.

In model selection and parametric estimation problems, (1) is referred to as the entropy loss. Among the parametric models, the one that minimizes the expected loss $E[K(F : F^* | \theta)]$ is selected. The entropy loss has been used with frequentist and Bayesian risk functions in model selection and parametric estimation problems; see Soofi (1997) and references therein.

In general, estimation of (1) directly is formidable. Akaike (1974) observed that decomposing the log-ratio in (1) gives

$$K(F : F^* | \theta) = -E_f[\log f^*(X|\theta)] - H(F), \quad (2)$$

where

$$H(F) = - \int \log f(x) dF(x) \quad (3)$$

is the entropy of F . Akaike information criteria (AIC) seeks the model $F^*(X|\theta)$ that minimizes (2) among a set of models. Since the entropy of the data-generating distribution $H(F)$ is free from the parameter θ , the second term in (2) is ignored in the derivation of the AIC criteria for model selection. The AIC-type measures are derived by minimizing the first term in (2) using the sample average of the likelihood function. Consequently, the AIC-type measures provide criteria for model comparison purposes only, and do not provide information diagnostics about the model fit. Spiegelhalter et al. (2002) developed a generalization of the AIC for hierarchical Bayesian models using the posterior deviance concept proposed by Dempster (1974). The authors referred to this model comparison criterion as the Deviance Information Criterion (DIC) and showed that DIC is asymptotically equivalent to AIC for nonhierarchical models.

For developing information indices that assess whether the unknown distribution $F(x)$ can be satisfactorily approximated by a parametric model $F^*(x|\theta)$, estimation of the entropy integral (3) plays the pivotal role. In the discrete case, the problem leads to the quantities related to sample proportion and log-linear and logit analysis. When F is absolutely continuous, information indices of fit and distributional tests require a nonparametric entropy estimate and an estimate for the entropy of a parametric model for F . This line of research began with Vasicek (1976) and continues to date. See for example, Vasicek (1976), Dudewicz and Van Der Meulen (1981), Gokhale (1983), Arizono and Ohta (1989), Ebrahimi et al. (1992), Soofi et al. (1995), DeWaal (1996), Ebrahimi (1997, 1998, 2001), Mazzuchi et al. (2000), Mudholkar and Tian (2002), Taufer (2002), Inverardi (2003), Park and Park (2003), Park (1999, 2005), and Choi and Kim (2006). With the exception of Mazzuchi et al. (2000), all of these articles use frequentist procedures for nonparametric estimation of entropy. Several frequentist procedures for estimation of entropy are available in the literature, see Dudewicz and Van Der Meulen (1987), Joe (1989), Hall and Morton (1993), Ebrahimi et al. (1994), Beirlant et al. (1997), Kraskov et al. (2004), and references therein.

Bayesian estimation of entropy has not received much attention. Gill and Joanes (1979) addressed Bayesian estimation of discrete entropy for the frequency data. Mazzuchi et al. (2000) proposed a computational procedure for Bayesian inference about the entropy and an information index of fit. The procedure was shown to be successful in identifying the correct model when fitting various models to a simulated example. Recently, Dadpay et al. (2007) have used a histogram-type entropy estimate for model fitting in generalized gamma family. Bayesian estimation of entropy is closely related to the notion of expected information in Bayesian analysis (Bernardo, 1979; Zellner, 1991) and to the notion of average entropy in the communication theory (Campbell, 1995).

For constructing information indices of fit, the parametric model $F^*(x|\theta)$ is selected based on the maximum entropy characterization of the densities of the parametric families. Consider the moment class of distributions

$$\Omega_{\theta} = \{F(x|\theta) : E_F[T_j(X)|\theta] = \theta_j, j = 1, \dots, J\}, \quad (4)$$

where $T_j(X)$ are integrable functions with respect to the density and $\theta = (\theta_1, \dots, \theta_J)$ is a vector of moment parameters. The maximum entropy (ME) model in the moment class (4), if it exists, has the density in the following exponential form:

$$f^*(x|\theta) = C(\eta)e^{-\eta_1 T_1(x) - \dots - \eta_J T_J(x)}, \quad (5)$$

where the model parameters $\eta = (\eta_1, \dots, \eta_J)$ are the Lagrange multipliers in the ME procedure with

$$\eta_j = \eta_j(\boldsymbol{\theta}), \quad j = 1, \dots, J, \quad (6)$$

and $C(\boldsymbol{\eta})$ is the normalizing constant.

The entropy of the ME model is given by

$$H[F^*(x | \boldsymbol{\theta})] = -\log C(\boldsymbol{\eta}) + \eta_1 \theta_1 + \dots + \eta_J \theta_J. \quad (7)$$

Soofi et al. (1995) showed that if $F \in \Omega_\theta$ and F^* is the ME model in Ω_θ , then the first term in (2) becomes the entropy of F^* and

$$K(F : F^* | \boldsymbol{\theta}) = H[F^*(x | \boldsymbol{\theta})] - H[F(x | \boldsymbol{\theta})]. \quad (8)$$

The first term is the entropy of the parametric ME model (5) and the second term is the entropy of a distribution which is unknown other than the density is a member of a general moment class, $F \in \Omega_\theta$. This equality defines the information distinguishability (ID) between distributions in Ω_θ . ID statistics are obtained by estimating (8).

The relation (8) reduces the problem of estimating $K(F : F^* | \boldsymbol{\theta})$ to the problem of estimating the two entropies shown in Eq. (8). Many known parametric families of distributions are in the form of Eq. (5) and therefore are ME subject to specific forms of moment constraints. For a parametric model, one may easily identify the moment class Ω_θ by writing the density in the exponential form (5); see, e.g., Soofi et al. (1995).

Two main issues remain: estimation of the entropy of the unknown distribution $H[F(x | \boldsymbol{\theta})]$ and maintaining the non-negativity of the estimate of (8).

This article develops a class of entropy estimates and provides a procedure for Bayesian inference on the entropy and a fit index. We define a quantized approximation of (3), which for the continuous case converges to the entropy integral. The quantized entropy provides a general representation for several existing entropy estimates, including Vasicek's sample entropy, which has been the main vehicle for the frequentist entropy-based distributional tests, and the existing Bayes entropy estimates (Dadpay et al., 2007; Gill and Joanes, 1979; Mazzuchi et al., 2000). We then derive a Bayes estimate of the quantized entropy, based on the Dirichlet process posterior for F , with the ME model as the prior expectation $E(F) = F^*$. We refer to this inferential procedure as the Maximum Entropy Dirichlet (MED). We explore the large sample properties of the Bayes estimates of entropy and the fit index.

For inference about the fit of the model F^* , we give an approximation of the posterior entropy loss, $K(F : F^* | \boldsymbol{\theta})$. The normalized estimated

average entropy loss provides a Bayesian Information Distinguishability (BID) index of fit for the model. Inference about the fit is based on comparing the prior and posterior distributions of the normalized $K(F : F^* | \theta)$. As byproducts, the MED also provides priors and posteriors for the moment parameters θ and the model parameters η .

Section 2 presents the quantized entropy and the Bayes estimate of entropy for continuous case. Section 3 relates the entropy estimate to some measures for the discrete case. Section 4 presents estimation of the Kullback–Leibler function. Section 5 reports some results of a simulation study. Section 6 gives some concluding remarks. Technical details are given in an Appendix.

2. THE CONTINUOUS CASE

For a real-valued random variable X with distribution function F , the probability integral transform gives a uniform random variable $U = F(X)$ over the unit interval and the quantile function is defined as $\xi = Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Parzen (2004) gives an insightful discussion of quantile functions and the useful roles that they play in statistical data modeling, including quantile formulas for mean and variance. Quantile formulas for sample moments use order statistics. Given a set of observations x_1, \dots, x_n from F , the sample quantiles are defined by the order statistics $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ with $\widehat{F}_i = \widehat{P}(X \leq y_{(i)}) = \frac{i}{n}$, see Parzen (2004).

The quantile formula for entropy (3) is

$$H(F) = \int_0^1 \log \left[\frac{d}{du} Q(u) \right] du. \quad (9)$$

Vasicek (1976) noted this entropy representation and defined a sample entropy based on order statistics as follows. At each sample point $(y_{(i)}, \frac{i}{n})$, the derivative in Eq. (9) is estimated by

$$s_i(m, n) = \frac{y_{(i+m)} - y_{(i-m)}}{\widehat{F}_{(i+m)} - \widehat{F}_{(i-m)}} = \frac{y_{(i+m)} - y_{(i-m)}}{2m/n}, \quad (10)$$

where $m \in \{1, 2, \dots, \leq \frac{n}{2}\}$ is an estimation window size. Vasicek's sample entropy is defined by the average of logarithm of $s_i(m, n)$ defined in (10),

$$H_v(m, n) = \frac{1}{n} \sum_{i=1}^n \log \left[\frac{n}{2m} (y_{(i+m)} - y_{(i-m)}) \right], \quad (11)$$

where $y_{i-m} = y_1$ for $i \leq m$ and $y_{i+m} = y_n$ for $i \geq n - m$.

We generalize Vasicek's sample entropy by a quantized entropy.

2.1. Quantized Entropy

Let $\xi_k = Q(u_k), k = 0, 1, \dots, q$ be a set of quantiles of F and let $F_k = F(\xi_k)$. Each set of quantiles defines a partition of the real line \mathfrak{R} ,

$$-\infty < \xi_0 < \xi_1 < \dots < \xi_q < \infty. \tag{12}$$

At each point (ξ_k, F_k) , we estimate the derivative in Eq. (9) by

$$s_k(m, q) = \frac{\Delta \xi_{m,k}}{\Delta F_{m,k}}, \tag{13}$$

where $m \in \{0, 1, 2, \dots, \leq \frac{q}{2}\}$ is an estimation window size, $\Delta \xi_{m,k}$ is spacing of order $2m$ defined by

$$\Delta \xi_{m,k} = \begin{cases} \xi_k - \xi_{k-1} \equiv \Delta \xi_k & \text{for } m = 0 \\ \xi_{k+m} - \xi_{k-m} = \sum_{\ell=-m}^m \Delta \xi_{k+\ell} & \text{for } m \geq 1, \end{cases} \tag{14}$$

and

$$\Delta F_{m,k} = \begin{cases} F_k - F_{k-1} \equiv \Delta F_k & \text{for } m = 0 \\ F_{k+m} - F_{k-m} = \sum_{\ell=-m}^m \Delta F_{k+\ell} & \text{for } m \geq 1, \end{cases} \tag{15}$$

with $\xi_{k-m} = \xi_0, k < m$ and $\xi_{k+m} = \xi_q, k > q - m$ are set such that $F_0 < \epsilon_0$ and $F_q < 1 - \epsilon_q$ for some small values ϵ_0 and ϵ_q .

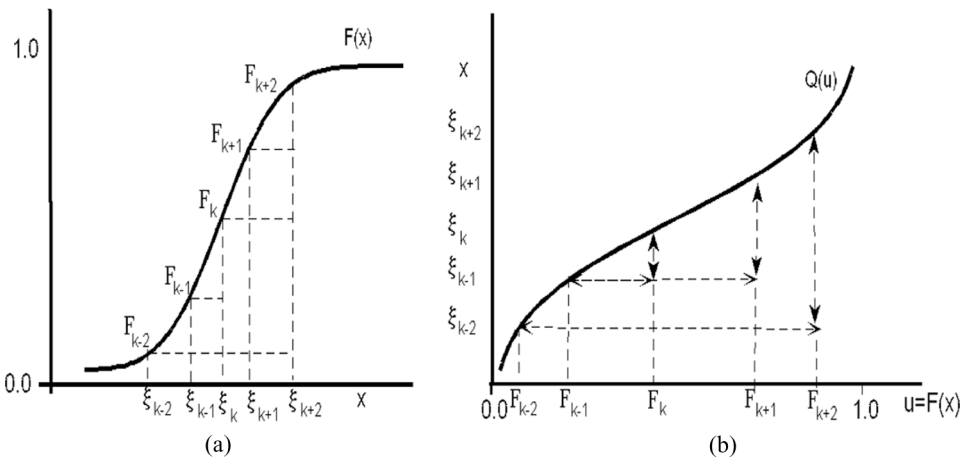


FIGURE 1 (a) Probability distribution function; (b) Quantile function.

Figure 1 depicts estimation of the derivative of the quantile function $\frac{dQ(u)}{du}$. Figure 1(a) shows five quantiles $\zeta_{k-2}, \zeta_{k-1}, \zeta_k, \zeta_{k+1}, \zeta_{k+2}$ and the probabilities given by the distribution function $F(x)$. Figure 1(b) shows the quantile function $\xi = Q(u)$, which is the mirror image of the $u = F(x)$ on the rotated axis. At point (F_k, ζ_k) , the derivative $\frac{dQ(u)}{du}$ may be estimated by any of the slopes computed by the following ratios of the spacings to the corresponding probability increments:

- (a) Ratio of the spacings of order one to the corresponding probability increments. $s_k(0, q) = \frac{\zeta_k - \zeta_{k-1}}{F_k - F_{k-1}}$;
- (b) Ratio of the spacings of order two to the corresponding probability increments. $s_k(1, q) = \frac{\zeta_{k+1} - \zeta_{k-1}}{F_{k+1} - F_{k-1}}$;
- (c) Ratio of the spacings of order four to the corresponding probability increments. $s_k(2, q) = \frac{\zeta_{k+2} - \zeta_{k-2}}{F_{k+1} - F_{k-1}}$.

Other plausible alternatives include $\frac{\zeta_{k+1} - \zeta_k}{F_{k+1} - F_k}$ and a combination of $s_k(m, q)$, $m = 1, 2$.

The roles of spacing $\Delta \xi_{m,k}$ and window size m are to smooth the slope estimate (13). With $m \geq 1$ the derivative is estimated based on the spacings of order $2m$ and the corresponding probability increments. Due to the relationship between the single and higher order spacings and probability increments shown in (14) and (15), the case of $m \geq 1$ provides a “moving average” type estimate for $\frac{dQ(u)}{du}$ in the form of

$$s_k(m, q) = \frac{\overline{\Delta \xi}_k}{\overline{\Delta F}_k}, \quad m > 0,$$

where $\overline{\Delta \xi}_k = \sum_{\ell=-m}^m \frac{\Delta \xi_{k+\ell}}{2m}$ and $\overline{\Delta F}_k = \sum_{\ell=-m}^m \frac{\Delta F_{k+\ell}}{2m}$ are the moving averages of spacings of order one and the corresponding probability increments defined by $m \geq 1$. If the spacings of order one are all equal, then the case of $m > 0$ is simply the moving average of the case of $m = 0$, i.e., $s_k(m, q) = \bar{s}_k(0, q)$ ¹.

The quantized entropy is defined by the average of the logarithm of the slopes (13)

$$H_{m,q}(F) = - \sum_{k=1}^q \Delta F_k \log \frac{\Delta F_{m,k}}{\Delta \xi_{m,k}}. \tag{16}$$

For a distribution with an absolutely continuous density, $H_{m,q}(F)$ provides a quantized approximation for the entropy integral (3). Note that

¹Interpretation of this special case is due to a referee

$H_{m,q}(F)$ has the scale property of the differential entropy $H(F)$. That is for a constant a , $H_{m,q}(F_{aX}) = H_{m,q}(F_X) + \log|a|$.

The quantized entropy $H_{m,q}(F)$ is a general representation that includes several entropy estimates proposed in the literature. For $m = 0$, the quantized entropy (16) gives a modification of the quantized approximation of the differential entropy (3) by the discrete entropy given by Cover and Thomas (1991). This case was used in Mazzuchi et al. (2000). The modification resolves the convergence problem of the quantized approximation of Cover and Thomas (1991). The following lemma encapsulates this property.

Lemma 1. *Let F be a distribution with an absolutely continuous density f . Then,*

$$H_{m,q}(F) \rightarrow H(F), \quad \text{as } q \rightarrow \infty, \quad \Delta \xi_{m,k} \rightarrow 0 \quad \forall k.$$

Proofs of this and other results are given in the Appendix.

When the partition (12) is defined in terms of the order statistics, $\xi_k = y_i, k = i = 1, \dots, n$, then $H_{m,n}(F)$ with $\xi_1 = y_1, \dots, \xi_q = y_n$, and the empirical distribution $\Delta F_n = \Delta \widehat{F}_n = 1/n$ gives Vasicek's sample entropy $H_v(m, n)$. Ebrahimi et al. (1994) observed that Eq. (10) does not define correct measures of the slope when $i \leq m$ or $i \geq n - m + 1$ and provided two modifications of $H_v(m, n)$ which are asymptotically equivalent to $H_v(m, n)$. The quantized entropy (16) includes both of these modifications as specific cases.

Hall and Morton (1993) proposed estimating $H(f)$ by the entropy of a histogram,

$$H_q^h(F) = - \sum_{k=1}^q \Delta \widehat{F}_k \log \frac{\Delta \widehat{F}_k}{h},$$

where $\Delta \widehat{F}_k = \frac{n_k}{n}, k = 1, \dots, q$ are the relative frequencies in the binwidth $h = \Delta \xi_{m,k}$. For $m = 0$ and $\Delta \xi_k = h$, $H_{0,q}(F)$ is a generalization of the histogram entropy and gives $H_q^h(F)$ when $\Delta F_k = \Delta \widehat{F}_k, k = 1, \dots, q$.

Another useful partition is given by the Dirichlet tessellation (Bowyer, 1981) defined by

$$\mathcal{T}_k = \{x : |x - \xi_k| < |x - \xi_j|, \forall j \neq k\}.$$

In this case each data point $x_i, i = 1, \dots, n$ is placed in an interval $(\xi_i, \xi_{i+1}), i = 0, \dots, n - 1$.

The boundaries are given by the data midpoints

$$\xi_i = \frac{1}{2}(y_{(i)} + y_{(i+1)}), \quad i = 1, \dots, n - 1, \tag{17}$$

where $y_{(1)} \leq \dots \leq y_{(n)}$ are the order statistics; ξ_0 and ξ_n are defined as above.

Theil (1980) used the midpoints (17) for derivation of a mass and mean preserving ME (MMME) density estimate f_n^* , which is piecewise uniform over the intervals $[\xi_i, \xi_{i+1}]$, $i = 1, \dots, n-1$ and exponential over $(-\infty, \xi_1]$, and $[\xi_n, \infty)$. Its entropy is given by

$$H(F_n^*) = - \sum_{i=1}^n \Delta \widehat{F}_i \log \frac{\Delta \widehat{F}_i}{\Delta \xi_i} + \frac{2}{n} (1 - \log 2).$$

This entropy is used as a nonparametric entropy (Dudewicz and Van Der Meulen, 1987; Soofi and Retzer, 2002) and it is a limiting case of the quantized entropy (16) with $F = F^*$, $m = 1$, and $q = n$ as $n \rightarrow \infty$.

In case of histogram partition, where $h = \Delta \xi_{m,k}$, $\forall k$, and $m = 0$, smoothing is done only through the binwidth h . Whereas in the case of a partition defined by order statistics (e.g., Vasicek's sample entropy and Dirichlet tessellation) smoothing is done through the window size. The choice of $m \geq 1$ for Vasicek's sample entropy and its modifications have been studied via simulation in terms of frequentist criteria such as bias and mean squared error of estimation and power of the distributional tests (see, e.g., Ebrahimi et al., 1992, 1994; Mudholkar and Tian, 2002; Park and Park, 2003). Choice of m for the quantized entropy requires simulation study in the context of its applications to various statistical problems. Section 5 reports some results for Bayesian estimation of entropy and inference about the fit.

2.2. Bayes Estimate of Entropy

Given a sample $\mathbf{x} = (x_1, \dots, x_n)$ from unknown F , the Bayes entropy estimate is defined by the mean of a posterior distribution of the quantized entropy (16),

$$\widetilde{H}_{m,q}(F) \equiv E[H_{m,q}(F) | \mathbf{x}].$$

We use the Dirichlet process prior (Ferguson, 1973) for the unknown F ,

$$F(\xi_k) | \mathcal{B}, F^* \sim \mathcal{D}(\mathcal{B}, F^*), \quad (18)$$

where F^* is a prior mean of F and $\mathcal{B} > 0$ is the strength of belief parameter. We use F^* , the ME distribution in Ω_θ , as the best guess distribution of the Dirichlet prior. In this context, \mathcal{B} is the degree of belief in the ME distribution F^* and we refer to $\mathcal{D}(\mathcal{B}, F^*)$ as the *maximum entropy Dirichlet (MED)* prior.

For any partition (12) of \mathfrak{R} , the increments ΔF_k , $k = 1, \dots, q$ have the Dirichlet distribution

$$\pi(\Delta F) \propto (\Delta F_1)^{\alpha(\xi_1)-1} (\Delta F_2)^{\alpha(\xi_2)-\alpha(\xi_1)-1} \dots (\Delta F_q)^{\mathcal{B}-\alpha(\xi_{q-1})-1}. \tag{19}$$

The parameters of the Dirichlet distribution (19) are given by $\alpha(\xi_k) \equiv \alpha((-\infty, \xi_k])$, a measurable function defined over \mathfrak{R} such that $\lim_{\xi \rightarrow \infty} \alpha(\xi) = \mathcal{B}$, and

$$E[F(\xi_k)] = \frac{\alpha(\xi_k)}{\mathcal{B}} = F^*(\xi_k).$$

The prior variance of $F(\xi)$ is given by

$$V[F(\xi) | F^*, \mathcal{B}] = \frac{\alpha(\xi)[\mathcal{B} - \alpha(\xi)]}{\mathcal{B}^2(\mathcal{B} + 1)} = \frac{F^*(\xi)[1 - F^*(\xi)]}{\mathcal{B} + 1},$$

reinforcing the notion of F^* being the best guess distribution and \mathcal{B} being the strength of belief parameter.

It is well-known that the posterior distribution of F is also a Dirichlet process,

$$F(\xi) | F^*, \mathcal{B}, x \sim \mathcal{D}(\tilde{\mathcal{B}}, \tilde{F}), \tag{20}$$

with the parameters updated by the data by

$$\tilde{\mathcal{B}} = \mathcal{B} + n, \quad \text{and} \quad \tilde{\alpha}(\xi_k) = \alpha(\xi_k) + \sum_{i=1}^q \delta[x_i \leq \xi_k], \tag{21}$$

where $\delta[\cdot]$ is the indicator function of the set.

The posterior mean of F_k is given by

$$\tilde{F}_k = E[F(\xi_k) | F^*, \mathcal{B}, x] = \frac{\mathcal{B}}{\tilde{\mathcal{B}}} F^*(\xi_k) + \frac{n}{\tilde{\mathcal{B}}} \hat{F}(\xi_k), \tag{22}$$

where $\hat{F}(\xi_k)$ is an empirical distribution.

The natural choice for \hat{F} is the empirical distribution (piecewise uniform). For the Dirichlet tessellation with $m = 0$, the empiric distribution function of order one (piecewise uniform over intervals $\xi_{k-1} < x \leq \xi_{k+1}$, $k = 1, \dots, n$) developed by Dudewicz and Van Der Meulen (1987) and the cumulative distribution of Theil's MMME (piecewise uniform over intervals $\xi_{k-1} < x \leq \xi_{k+1}$, $k = 1, \dots, n$ with exponential tails) are also applicable. For Dirichlet tessellation with $m \geq 1$ the higher order empiric distribution developed by Dudewicz and Van Der Meulen (1987) and the distribution function associated with Vasicek's sample entropy developed by Park and Park (2003) are applicable.

Theorem 1. For each given partition (12) with the Dirichlet prior (19) the following results hold:

(i) The posterior average quantized entropy is

$$\tilde{H}_{m,q}(F) = \psi(\tilde{\mathcal{B}} + 1) - \sum_{k=1}^q \Delta \tilde{F}_k [\psi(\tilde{\mathcal{B}} \Delta \tilde{F}_{m,k} + 1) - \log \Delta \xi_{m,k}], \quad (23)$$

where $\psi(z) = \frac{d \log \Gamma(z)}{dz}$ is the digamma function;

(ii) For large $\tilde{\mathcal{B}}$ such that $\frac{\Delta \tilde{F}_k}{\tilde{\mathcal{B}} \Delta \tilde{F}_{m,k}} \approx 0$,

$$\tilde{H}_{m,q}(F) \approx - \sum_{k=1}^q \Delta \tilde{F}_k \log \frac{\Delta \tilde{F}_{m,k}}{\Delta \xi_{m,k}}. \quad (24)$$

Since the Dirichlet prior (19) requires specification of a parametric first guess distribution F^* , for $\mathcal{B} > 0$ the posterior mean $\tilde{H}_{m,q}(F)$ is a *semiparametric Bayes entropy estimate* under quadratic loss. For the case of the improper prior $\mathcal{B} = 0$, or when $n \rightarrow \infty$, the posterior mean $\tilde{H}_{m,q}(F)$ is a *nonparametric Bayes entropy estimate* under quadratic loss.

In practice, the posterior mean will be computed by the average of a large number of replications of posterior Dirichlet vectors generated from (20) and computing (16) for each run, and then taking the average. By the law of large numbers, the simulated Bayes estimate should be close to (23). The difference between the average of simulated results and (23) can be used assessing the simulation accuracy.

Next result gives the consistency of the Bayes entropy estimate for the histogram partition.

Theorem 2. Let F be a distribution with an absolutely continuous density f . Then $\tilde{H}_{m,q}(F)$ based on a histogram type partition is consistent.

In the case of Dirichlet tessellation partition, $q = n$ and $\tilde{H}_{m,q}(F) = \tilde{H}_{m,n}(F)$. Note that use of the Dirichlet tessellation partition circumvents the empty cell problem that usually is encountered in applications of the Dirichlet process prior. Next result gives the consistency of the Bayes entropy estimate for the Dirichlet tessellation partition.

Theorem 3. Let F be a distribution with an absolutely continuous density f . Let (12) be the partition with ξ_i defined by the Dirichlet tessellation (17). Then as $n \rightarrow \infty$, $m \rightarrow \infty$, and $\frac{m}{n} \rightarrow 0$,

$$\tilde{H}_{m,n}(F) \xrightarrow{p} H_v(m, n) \xrightarrow{p} H(F),$$

where $H_v(m, n)$ is the Vasicek's entropy estimate.

According to Theorems 2 and 3, consistency may be achieved by smoothing, either through the binwidth h or through increasing the window size m . Consistency can also be proved for the case of $m = 0$, however, under a very stringent requirement of $\Delta\xi_i \rightarrow 0 \forall i = 1, \dots, n$, which is hardly implementable. We therefore present the consistency for the case of $m = 0$ as a remark.

Remark 1. Let F be a distribution with an absolutely continuous density f . Let (12) be the partition with ξ_i defined by the Dirichlet tessellation (17). Then as $\Delta\xi_i \rightarrow 0 \forall i = 1, \dots, n$, $\tilde{H}_{0,n}(F) \xrightarrow{p} H(F)$.

We also should point out that the posterior variance of quantized entropy is available. Since derivation is tedious and the variance expression is messy, it is not reported here. It can be shown that the posterior variance goes to zero as $n \rightarrow \infty$.

3. THE DISCRETE CASE

We have discussed the problem when F has a continuous density. However, with some modifications, the procedure is applicable when F is discrete or a distribution over q categories. Let

$$H(\Delta F) = - \sum_{k=1}^q \Delta F_k \log \Delta F_k, \tag{25}$$

where $\Delta F_k = f_k$ is probability assigned by F to the k th category indexed arbitrarily, $k = 1, \dots, q$. Then under the Dirichlet prior (19) for F , the posterior average discrete entropy is

$$\tilde{H}(\Delta F) = \psi(\tilde{\mathcal{B}} + 1) - \sum_{k=1}^q \Delta \tilde{F}_k [\psi(\tilde{\mathcal{B}} \Delta \tilde{F}_k + 1)]. \tag{26}$$

This is obtained by using part (ii) of Lemma 2 of the Appendix with $u = \tilde{\mathcal{B}} \Delta \tilde{F}_k$ and $v = \tilde{\mathcal{B}}(1 - \Delta \tilde{F}_k)$.

Specific cases of (26) have been used previously. Gill and Joanes (1979) used the symmetric Dirichlet prior distribution

$$\pi(f_1, \dots, f_q) \propto f_1^{\rho-1} \dots f_q^{\rho-1}, \quad \rho > 0 \tag{27}$$

and found the posterior mean of $H(f)$. This prior reduces to the uniform prior when $\rho = 1$ and to the Jeffreys invariant prior when $\rho = 1/2$.

Campbell (1995) obtained simple expressions for the prior average entropy and the prior entropy variance using (27). The following theorem characterizes the results based on (27) in terms of the Dirichlet prior (19).

Theorem 4. *Let F be a distribution with probability mass function f_1, \dots, f_q over q points $y_1 < \dots < y_q$, where the order is arbitrary for the categorical case. Then the average entropy based on the symmetric Dirichlet prior (27) is given by (26) if and only if the prior guess for the Dirichlet prior (19) is uniform, $F^*(y_k) = k/q$, $k = 1, \dots, q$.*

Let $\Delta\hat{F}_k = n_k/n$, where $n_k = \sum_{i=1}^q \delta[x_i = y_k]$. Then with the uniform best guess, the posterior average entropy (26) gives the Bayes estimate of the discrete entropy obtained by Gill and Joanes (1979).

4. INFORMATION INDEX OF FIT

Entropy-based tests of distributional hypothesis $F = F^*$ are derived either based on difference between two entropies $H(F^*) - H(F)$ or based on the Kullback–Leibler information $K(F : F^*)$. The entropy difference tests are constructed using a parametric entropy estimate $H[F^*(x | \hat{\eta})]$, where $\hat{\eta}$ is an estimate of the model parameter and a nonparametric estimate of $H(F)$ (usually, Vasicek’s sample entropy $H_v(m, n)$). The Kullback–Leibler tests are also constructed using a parametric and a nonparametric entropy estimate in the right-hand-side of (8). Thus, both types of tests are constructed by entropy difference, however often without imposing the constraints (4) on F and F^* . When F is not constrained to be in $\Omega_{\hat{\theta}}$ in which $F^*(x | \hat{\theta})$ is the ME model, the entropy difference does not measure disparity between F and F^* . Moreover, without such constraint on F , the entropy difference in (8) may produce a negative estimate of $K(F : F^* | \theta)$. Soofi et al. (1995) and Park and Park (2003) developed distributional fit indices and tests with $F, F^* \in \Omega_{\hat{\theta}}$ ensuring non-negativity of $K(F : F^* | \hat{\theta})$.

In order to ensure the non-negativity of the Bayes estimate of $K(F : F^* | \theta)$, we estimate the parameters of the ME model (hyperparameters of Dirichlet prior) by the moments of the quantized F . Quantized approximations of the moments are obtained by

$$\theta_{q,j} = \sum_{k=1}^q T_j(\bar{\zeta}_{k,m})(\Delta F_k), \quad j = 1, \dots, J, \tag{28}$$

where

$$\bar{\zeta}_{k,m} = \begin{cases} \frac{\zeta_k + \zeta_{k-1}}{2} & \text{for } m = 0 \\ \frac{\zeta_{k-m} + \zeta_{k+m}}{2} & \text{for } m \geq 1. \end{cases}$$

The quantized moments (28) are approximations of the moments in (4); $\theta_{q,j} = \theta_j(\Delta F) \approx \theta_j(F) = \theta_j$. Use of (28) allows to compute samples

for the vector of quantized moments θ_q from the Dirichlet prior and posterior samples for F and construct prior and posterior for θ . Then the priors and posteriors for the model parameters η_j are found by $\eta_{q,j} = \eta_j(\theta_q)$.

The prior and posterior distributions of $H(F^* | \theta)$ are obtained using $\theta_{q,j}$ and $\eta_{q,j}$ in (7), which gives

$$H[F^*(x | \theta_q)] = -\log C(\eta_q) - \sum_{j=1}^J \eta_{q,j} \theta_{q,j}.$$

We then use the maximum entropy $H(F^* | \theta_q)$ and the quantized entropy $H_{m,q}(F)$ in (8) and obtain prior and posterior distributions for $K(F : F^* | \theta)$ by

$$K_{m,q}(F : F^*) = H(F^* | \theta_q) - H_{m,q}(F).$$

Since $P[H(F^* | \theta_q) \geq H_{m,q}(F)] = 1$, we have $P[K_{m,q}(F : F^*) \geq 0] = 1$. Note that $K_{m,q}(F : F^*)$ is scale invariant.

Using the posterior mean of the entropy of the ME model $\tilde{H}[F^*(x | \theta_q)]$ and the Bayes estimate of entropy (23) in (8) gives the posterior mean of $K_{m,q}$

$$\begin{aligned} \tilde{K}_{m,q}(F : F^*) &= \tilde{H}[F^*(x | \theta_q)] + \psi(\tilde{\mathcal{B}} + 1) \\ &\quad - \sum_{k=1}^q \Delta \tilde{F}_k [\psi(\tilde{\mathcal{B}} \Delta \tilde{F}_{m,k} + 1) - \log \Delta \xi_{m,k}] \\ &\approx \tilde{H}[F^*(x | \theta_q)] - \sum_{k=1}^q \Delta \tilde{F}_k \log \frac{\Delta \tilde{F}_{m,k}}{\Delta \xi_{m,k}}. \end{aligned} \tag{29}$$

Thus, $\tilde{K}_{m,q}(F : F^*)$ is an approximate Bayes estimate of $K(F : F^* | \theta)$ under quadratic loss. This Bayes estimate is an approximation of the expected entropy loss of estimating F by $F^*(x | \theta_q)$ and provides an *ID index* for assessing the fit of the parametric model F^* .

A normalized ID index is given by

$$\begin{aligned} ID_{m,q}(F : F^*) &= 1 - \exp[-K_{m,q}(F : F^*)] \\ &= 1 - \exp[H_{m,q}(F) - H(F^* | \theta_q)]. \end{aligned} \tag{30}$$

Note that $0 \leq ID_{m,q}(F : F^*) \leq 1$, and $ID_{m,q}(F : F^*) = 0$ if and only if $F(x | \theta) = F^*(x | \theta)$ with probability 1. The posterior mean $\tilde{ID}_{m,q}(F : F^*)$ of the $ID_{m,q}(F : F^*)$ index will be referred to as *Bayesian Information Discrimination (BID) index*. We compare the posterior and prior

distributions of the ID index for inference about the fit. If the posterior distribution of $ID_{m,q}(F : F^*)$ shifts to the left of the prior, then data provide support for the ME model. If the posterior shifts to the right of the prior or concentrates around the prior mean, then the data and model are not compatible.

Next result gives the consistency of the information index of fit for the histogram and tessellation partitions.

Theorem 5. *Let F^* be a distribution with an absolutely continuous density f^* . If the data is generated from F^* , then the following results hold:*

- (i) *For the histogram type partition $\tilde{K}_{m,q}(F : F^*) \xrightarrow{p} 0$;*
- (ii) *Let (12) be the partition with ξ_i defined by the Dirichlet tessellation (17). Then as $n \rightarrow \infty$, $m \rightarrow \infty$, and $\frac{m}{n} \rightarrow 0$, $\tilde{K}_{m,q}(F : F^*) \xrightarrow{p} 0$.*

Thus if F^* is not a suitable approximation for the true data-generating distribution F , then for large n , we can generally expect large values of $\tilde{K}_{m,q}$. The next remark gives the consistency for the case of $m = 0$ under the stringent requirement of $\Delta\xi_i \rightarrow 0 \forall i = 1, \dots, n$.

Remark 2. *Let F be a distribution with an absolutely continuous density f . Let (12) be the partition with ξ_i defined by the Dirichlet tessellation (17). Then as $\Delta\xi_i \rightarrow 0 \forall i = 1, \dots, n$, and as $n \rightarrow \infty$, $\tilde{K}_{0,q}(F : F^*) \xrightarrow{p} 0$.*

5. IMPLEMENTATION AND EXAMPLES

This section outlines the computational steps and reports results of a simulation study for the proposed Bayes entropy estimate and information index of fit.

5.1. Implementation

Computation of the entropy estimate and the fit index for the case of $m = 0$ is discussed in Mazzuchi et al. (2000). The procedure is adjusted for general m as follows.

1. Specification of the MED inputs:
 - (a) Specify the moment constraints (4) for which the likelihood function is the ME model $F^*(x | \theta)$; for a table of ME distributions see, e.g., Soofi et al. (1995);
 - (b) Set up the moment equations (28), the model parameter equations (6), and use the entropy expression (7) for $F^*(x | \theta)$; entropy expressions for many well-known parametric families of distributions are available, see, e.g., Ebrahimi et al. (1999);

- (c) Specify the degree of belief parameter \mathcal{B} , which reflect the uncertainty about the ME distribution $F^*(x|\theta)$ for the prior expected distribution;
- (d) Set the best-guess moment parameter θ^0 ; in the absence of prior values use the data moments;
- (e) Set ξ_0 and ξ_q as suitable percentiles of $F^*(x|\theta^0)$ and compute data midpoints ξ_i for the Dirichlet tessellation partition (or determine the histogram binwidth h).

2. Simulation of the MED priors:

- (a) Simulate n -dimensional Dirichlet vectors, $F^s = (\Delta F_1^s, \dots, \Delta F_n^s)$, $s = 1, \dots, S$, $S \gg n$, according to 1(a)–1(e) and obtain the Dirichlet prior (18) for F ;
- (b) For each Dirichlet vector F^s , $s = 1, \dots, S$, use (16) to compute $H_{m,q}^s(F)$ and obtain the prior distribution for the unknown entropy $H(F)$;
- (c) For each Dirichlet vector F^s , $s = 1, \dots, S$, use (28) to compute the quantized moment parameters θ_q^s and obtain the prior distributions of the moment parameters θ ;
- (d) For each vector of moment parameters θ_q^s , $s = 1, \dots, S$, use (6) to compute the model parameters η_q^s and obtain the prior distributions of η ;
- (e) For each vector of moments and vector of model parameters θ_q^s, η_q^s , $s = 1, \dots, S$, use the entropy expression (7) to compute the ME model entropy $H(F^*|\theta_q^s)$ and obtain the prior for $H(F^*|\theta)$;
- (f) Use each pair of $H_{m,q}^s(F)$ and $H(F^*|\theta_q^s)$, $s = 1, \dots, S$ in (30) to compute $ID_{m,q}^s(F : F^*; \theta_q)$ and obtain the prior distribution for the ID index;
- (g) The simulation accuracy can be checked as follows:
 - i. Compute the prior mean of the quantized entropy using (23) with $n = 0$ in (22) and compare it with the simulated mean $\overline{H}_{m,q}^s = S^{-1} \sum_{s=1}^S H_{m,q}^s$. If the two quantities are not satisfactorily close, increase the number of simulations S ;
 - ii. Compare the quantized moments with the corresponding moments of $F^*(x|\theta^0)$. If there is a large discrepancy, adjust ξ_q (h for histogram partition).

3. Simulation of the MED posteriors:

- (a) Update the Dirichlet prior parameters to obtain the posterior parameters of (21);

- (b) Simulate S posterior Dirichlet vectors using 1(a)–(d) with the updated Dirichlet parameters and obtain the posterior distribution of the quantized distribution F ;
- (c) Obtain the MED posterior distributions using the posterior Dirichlet vectors as in 2(a) and then follow 2(b)–(g).

5.2. Example

We present results for three sets of data generated from exponential, gamma, and lognormal distributions. For each set of data, we consider all three distributions as the candidate ME models. These results are typical and may be viewed as prototypes of MED analysis for these types of data and models.

Table 1 shows the distribution function $F^*(x|\theta)$ of each candidate model used as the best guess in the Dirichlet prior for unknown distribution F , the moment class (4) in which it is the ME model, and the expression for the entropy of the ME model. The table also shows the quantized moment equations (28) which link the ingredients of the MED procedure. We complete the MED prior specification for each model by setting $\mathcal{B} = 6$ to reflect a very weak degree of belief in the ME distribution.

We report on the results for the histogram partition and the Dirichlet tessellation for $n = 500$. For the histogram bin h , we started with a standard rule (Sturges’ rule $h = 1 + 3.322 \log(500)$), which gives approximately 10 classes. But then this was adjusted for accuracy in moment fitting. We used 12–15 classes in most cases. However, for lognormal alternative sometimes a large number of classes were needed in order to obtain moment estimates close to sample moments, and hence to prevent negative values for $K_{m,q}(F : F^*)$. Computation time on a desk top for histogram partition was about 10 seconds.

For the Dirichlet tessellation, we report the results for window sizes $m = 0, 1, 2, 5, 10$, which depict the general patterns in light of the

TABLE 1 Maximum entropy models used for simulated data

ME model	Exponential	Gamma	Lognormal
Distribution	$1 - e^{-\lambda x}$	$\int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t} dt$	$\int_0^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log t - \nu)^2} dt$
Moment class	$E(X) = \mu$	$\begin{cases} E(X) = \mu \\ E(\log X) = \nu \end{cases}$	$\begin{cases} E(\log X) = \nu \\ E(\log X)^2 = \tau \end{cases}$
Model entropy	$1 + \log \mu$	$\log \frac{\Gamma(\alpha)}{\lambda} + (1 - \alpha)\psi(\alpha) + \alpha$	$\frac{1}{2} \log(2\pi\sigma^2) + \nu$
Quantized moments	$\sum \Delta F_k \bar{\zeta}_{k,m} = \mu_q$	$\begin{cases} \sum \Delta F_k \bar{\zeta}_{k,m} = \mu_q \\ \sum \Delta F_k \log \bar{\zeta}_{k,m} = \nu_q \end{cases}$	$\begin{cases} \sum \Delta F_k \log \bar{\zeta}_{k,m} = \nu_q \\ \sum \Delta F_k (\log \bar{\zeta}_{k,m})^2 = \tau_q \end{cases}$

theoretical results. For the distributions under consideration here, the natural choice for $\zeta_0 = 0$. We set $\zeta_q = \zeta_{500}$ such that $F^*(\zeta_q) \geq .99$, adjusted to obtain moment estimates close to sample moments, and hence to prevent negative values of $K_{m,q}(F : F^*)$. Computation times on a desk top varied between 20–30 seconds for various values of m .

The exponential data were generated using $\lambda = 1$. Thus, the model entropy is $H(F^*) = 1$. The sample mean is $\bar{x} = 1.03$. The gamma data were generated using $\alpha = 2$ and $\lambda = 1$. Thus, the model entropy is $H(F^*) = 1.577$. The sample moments are $\bar{x} = 1.99$ and $\overline{\log x} = .41$, which give moment estimates (MLE) $\hat{\alpha} = 1.950$ and $\hat{\lambda} = .980$. The lognormal data were generated using $\nu = 0$ and $\sigma^2 = 1$. Thus, the model entropy is $H(F^*) = 1.419$. The sample moments are $\overline{\log x} = -0.012$ and $(\overline{\log x})^2 = 0.957$, which give MLE $\hat{\nu} = 1.950$ and $\hat{\sigma}^2 = .975$.

TABLE 2 Posterior means and standard deviations of entropies for exponential, gamma, and lognormal data

		Exponential $H = 1.000$		Gamma $H = 1.722$		Lognormal $H = 1.419$	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
<i>Maximum likelihood</i>							
	Model entropy \hat{H}^*	1.030	—	1.579	—	1.385	—
<i>Histogram</i>							
	Prior model entropy H^*	1.001	.348	1.478	.304	1.276	.437
	Posterior model entropy H^*	1.037	.040	1.607	.035	1.376	.052
	Prior quantized entropy $\tilde{H}_{m,q}$.659	.254	.952	.230	.780	.253
	Posterior quantized entropy $\tilde{H}_{m,q}$	1.005	.039	1.568	.033	1.350	.044
<i>Tessellation</i>							
$m = 0$	Prior model entropy H^*	.979	.371	1.480	.313	1.240	.449
	Posterior model entropy H^*	1.032	.042	1.605	.037	1.387	.054
	Prior quantized entropy $\tilde{H}_{m,q}$	-2.487	.529	-1.944	.499	-2.155	.615
	Posterior quantized entropy $\tilde{H}_{m,q}$.320	.061	.906	.054	.667	.069
$m = 1$	Prior model entropy H^*	.961	.362	1.454	.321	1.262	.444
	Posterior model entropy H^*	1.033	.041	1.604	.036	1.388	.055
	Prior quantized entropy $\tilde{H}_{m,q}$	-1.891	.507	-1.366	.468	-1.536	.587
	Posterior quantized entropy $\tilde{H}_{m,q}$.604	.055	1.172	.051	.951	.064
$m = 2$	Prior model entropy H^*	.948	.357	1.468	.302	1.253	.448
	Posterior model entropy H^*	1.031	.041	1.603	.036	1.386	.055
	Prior quantized entropy H^*	-1.304	.489	-0.751	.421	-.928	.576
	Posterior quantized entropy $\tilde{H}_{m,q}$.797	.051	1.355	.046	1.147	.062
$m = 5$	Prior model entropy H^*	.957	.372	1.478	.298	1.251	.453
	Posterior model entropy H^*	1.031	.041	1.604	.037	1.386	.055
	Prior quantized entropy $\tilde{H}_{m,q}$	-.566	.482	-.021	.422	-.198	.555
	Posterior quantized entropy $\tilde{H}_{m,q}$.942	.047	1.499	.042	1.299	.060
$m = 10$	Prior model entropy H^*	.962	.384	1.461	.309	1.258	.439
	Posterior model entropy $\tilde{H}_{m,q}$	1.032	.042	1.602	.035	1.384	.056
	Prior quantized entropy H^*	-.045	.490	.489	.402	.335	.536
	Posterior quantized entropy $\tilde{H}_{m,q}$	1.002	.047	1.556	.040	1.364	.058

Table 2 shows the results for entropy estimation. The maximum likelihood method gives estimate $\widehat{H}(F^*)$ for entropies of the parametric models. The MED method gives estimates of entropies of the parametric models $\widetilde{H}[F^*(x|\theta_q)]$ and a nonparametric estimate $\widetilde{H}_{m,q}(F)$ for each class of distributions where the model is ME. Table 2 shows the prior and posterior means and standard deviations for the model and quantized entropies. The table shows the following general patterns, typically found in the simulation runs. Estimates of model entropy given by MLE, posterior MED with histogram partition and tessellation for all m are all similar. For the exponential data, the model entropy estimates are close to the actual value. For gamma and lognormal data, the model entropies are underestimated by these methods. The quantized entropy estimates with histogram partition are close to the actual model entropies, thus verifying the result of Theorem 2. The quantized entropy estimates with tessellation partition move closer to the actual model entropies as m increases. For $m = 10$, the estimates are close to the actual model entropies, verifying the result of Theorem 3. The results for $m = 0$ are not satisfactory because the condition of $\Delta\xi_i \rightarrow 0, \forall i = 1, \dots, 500$ is violated; in general it is difficult to achieve this condition. Smoothing by $m > 0$ substantially improves the estimates.

Some general remarks are in order.

1. Entropy estimation procedures underestimate the entropy. This point has been noted in previous studies (e.g., Ebrahimi et al., 1994) and is observed for all procedures used in the present study, as noted in Table 2.
2. Table 2 shows substantial shifts from the prior means (highly negative) toward the actual model entropies for the MED procedures with histogram partition and tessellation for every m .
3. When the sample size is large and the prior is weak, by Theorems 2 and 3, the sample dominates the first guess model F^* , and this makes the MED a robust procedure for inference about entropy. Our simulation studies confirm this. Figure 2 shows the prior and posterior distributions of quantized entropy for gamma data when the gamma, exponential, and lognormal distributions described above are used as the best guess models in the prior. With $n = 500$ and $\mathcal{B} = 6$, the MED prior and posterior distributions are quite similar in all three cases. The factor that makes a difference is the information fit measure $\widetilde{K}_{m,q}(F : F^*)$ for these models is the parametric model entropy.

Table 3 shows the prior and posterior means and standard deviations of the normalized information fit index for three candidate models for the exponential data. Since exponential is gamma with shape parameter one (estimate based on exponential data is close to one), the values of

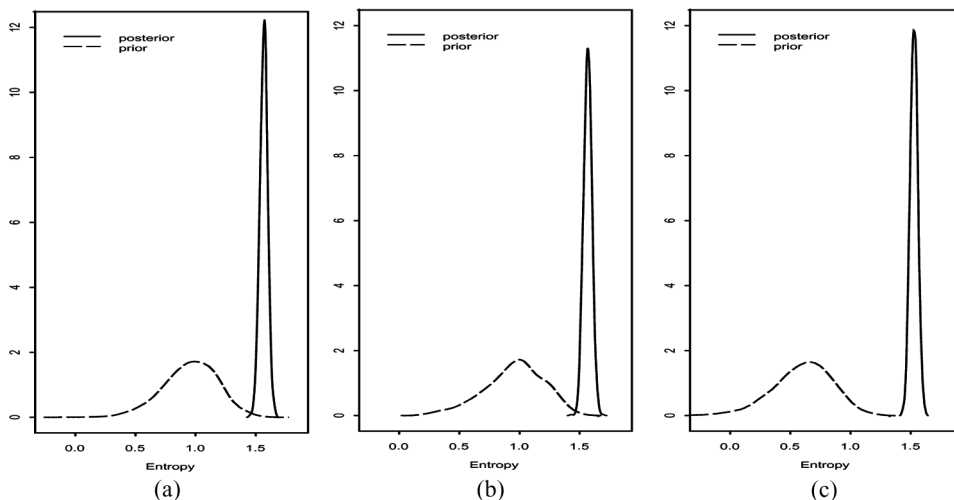


FIGURE 2 MED prior and posterior distributions of quantized entropy based on the histogram partition for gamma data with three best guess models for the prior. (a) Gamma model; (b) Exponential model; (c) Lognormal model.

posterior mean (BID) index is similar for the two models. The exponential BID is always lower than the lognormal BID. The consistency of the BID estimate is observed in the cases of histogram partition and tessellation partition with $m = 10$. In these cases, the BID indices of the exponential and gamma models are relatively much lower than the BID index of lognormal model.

TABLE 3 Prior and posterior means and standard deviations of fit indices for exponential data

		Candidate models					
		Exponential		Gamma		Lognormal	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
<i>Histogram</i>							
	Prior	.274	.136	.319	.152	.542	.152
	Posterior	.032	.006	.023	.008	.082	.013
<i>Tessellation</i>							
$m = 0$	Prior	.967	.012	.964	.014	.961	.016
	Posterior	.509	.021	.508	.021	.555	.021
$m = 1$	Prior	.939	.021	.934	.025	.929	.029
	Posterior	.349	.023	.347	.023	.406	.025
$m = 2$	Prior	.889	.034	.878	.047	.875	.040
	Posterior	.209	.022	.208	.022	.279	.026
$m = 5$	Prior	.773	.067	.755	.072	.745	.068
	Posterior	.085	.017	.084	.019	.164	.025
$m = 10$	Prior	.623	.100	.596	.108	.598	.096
	Posterior	.030	.015	.030	.015	.108	.024

Table 4 shows the prior and posterior means and standard deviations of the normalized information fit index for three candidate models for the gamma and lognormal data. For each data set, the BID for the true model is the lowest among the three candidate models. Again, for the histogram and tessellation with $m = 10$, differences between the BID indices of gamma and other two models are relatively substantial.

Figure 3 shows the MED prior and posterior distributions of the information fit index of the gamma and exponential models for the gamma data. We note that in all cases the posterior distributions shift toward zero and concentrate relative to the prior, and more so for the gamma model than for the exponential. That is, the fit index favors

TABLE 4 Prior and Posterior means and standard deviations of fit indices for gamma and lognormal data

		Candidate models					
		Exponential		Gamma		Lognormal	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Gamma data							
<i>Histogram</i>							
	Prior	.448	.144	.391	.139	.541	.138
	Posterior	.113	.014	.038	.008	.089	.013
<i>Tessellation</i>							
$m = 0$	Prior	.960	.021	.965	.012	.965	.013
	Posterior	.542	.020	.503	.021	.506	.021
$m = 1$	Prior	.930	.031	.937	.021	.936	.021
	Posterior	.403	.022	.350	.022	.355	.023
$m = 2$	Prior	.878	.046	.887	.032	.886	.033
	Posterior	.282	.022	.219	.021	.229	.022
$m = 5$	Prior	.765	.072	.769	.060	.765	.057
	Posterior	.175	.019	.099	.016	.125	.018
$m = 10$	Prior	.627	.097	.611	.092	.611	.087
	Posterior	.130	.018	.044	.013	.086	.013
Lognormal data							
<i>Histogram</i>							
	Prior	.169	.083	.339	.154	.361	.174
	Posterior	.118	.026	.105	.034	.026	.017
<i>Tessellation</i>							
$m = 0$	Prior	.963	.015	.960	.023	.964	.013
	Posterior	.553	.022	.552	.023	.513	.021
$m = 1$	Prior	.933	.026	.927	.032	.936	.023
	Posterior	.407	.025	.404	.026	.354	.024
$m = 2$	Prior	.878	.046	.887	.032	.886	.033
	Posterior	.282	.022	.219	.021	.229	.022
$m = 5$	Prior	.765	.066	.744	.079	.756	.066
	Posterior	.159	.029	.157	.027	.083	.019
$m = 10$	Prior	.619	.098	.588	.107	.591	.095
	Posterior	.097	.029	.094	.029	.020	.013

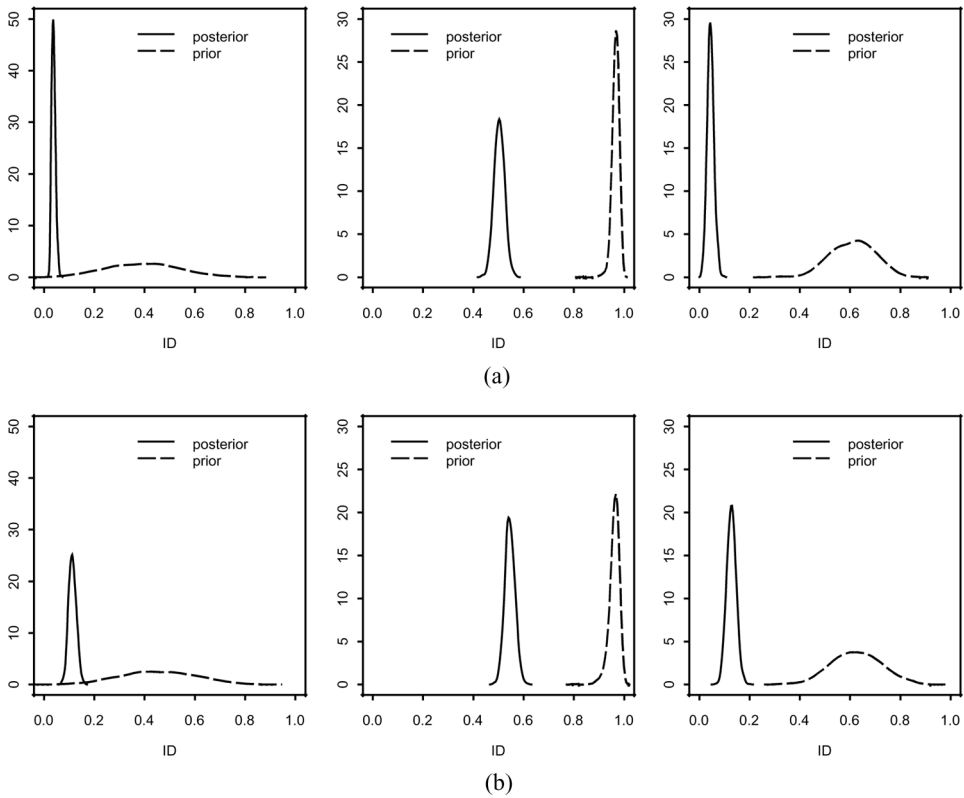


FIGURE 3 MED prior and posterior distributions of information index of fit for two alternative models for gamma data. (a) Gamma model; (b) Exponential model.

the correct model for any partition type and window size. The posterior distributions for the gamma based on the histogram (left panel) and tessellation partition with $m = 10$ (right panel) depict the consistency of MED for estimation of Kullback–Leibler information encapsulated in Theorem 5. The consistency for the case of $m = 0$ (middle panel) is not achieved due to the lack of the required condition $\Delta\xi \rightarrow 0, \forall i = 1, \dots$

Figure 4 shows the MED prior and posterior for the gamma model parameters based on the gamma data, which is produced through the quantized moment estimation. We note that the posterior distributions are almost symmetric and concentrate near the true values of the parameters.

6. CONCLUDING REMARKS

We have introduced a quantized entropy which provides a general representation of various versions of sample entropy measures found in the literature. We have developed a Bayesian alternative to the sampling

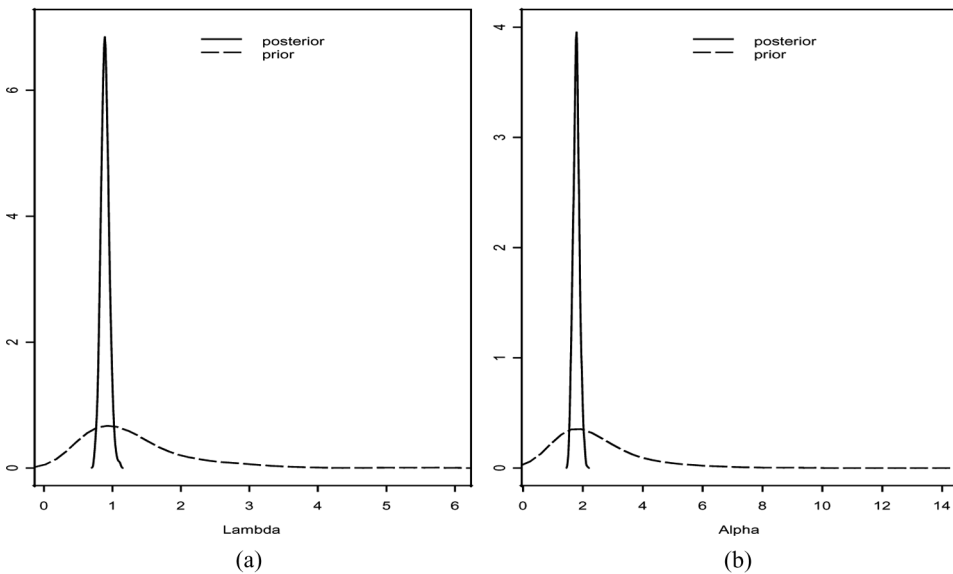


FIGURE 4 MED prior and posterior distributions of parameters of gamma distribution. (a) Lambda; (b) Alpha.

theory entropy-based approach for assessing the fit of parametric models. The fit is measured by an estimate of a Kullback–Leibler information. Prior and posterior distributions for the quantized entropy are obtained from a Dirichlet process prior for the unknown data generating distribution. We then showed how prior and posterior distributions for the information index can be obtained based on the maximum entropy characterization of the parametric model. The ME model is used as the expectation in the Dirichlet prior. In addition to providing Bayesian inference for the entropy and information index of fit, the MED procedure produces prior and posterior distributions for the model parameters and the moments.

We have derived the exact expression and a simple approximate formula for the posterior mean of the quantized entropy. The posterior mean provides a semiparametric Bayes estimate of entropy, which becomes nonparametric when the sample dominates the prior. The posterior variance is also available, but is not included in this article because the expression for the variance and its derivation are cumbersome. We have shown that the Bayes estimates of entropy and the Kullback–Leibler information possess appropriate consistency properties. The inference is computer-intensive. All prior and posterior distributions are found via simulations. The computation algorithm is outlined.

We have reported results of examples of some simulation runs for exponential, gamma, and lognormal data, where for each set of data, all three distributions were considered as the ME models. The findings,

which are prototypes of MED analysis for these types of data and models shed some light on the theoretical results. The examples indicated that MED is as good as the MLE for producing estimates of parametric model entropy. In addition, MED produces nonparametric estimate and Bayesian inference for entropy, which for large sample is robust against the best guess model in the Dirichlet prior. The proposed Kullback–Leibler information index, BID, favored the correct model in every case. The simulation results indicated that smoothing (through histogram or large window size) is important for consistency of the Bayes entropy estimate and the Kullback–Leibler information index. The case of $m = 0$, although selected the correct model in all three cases, it failed to verify its consistency results with a sample as large as $n = 500$ and a Dirichlet prior as weak as one signified by $\mathcal{B} = 6$.

Finally, two points are noteworthy regarding our formulation of the Dirichlet prior. First, in the MED prior, the initial best guess model is formulated systematically according to the maximum entropy formalism. Second, the MED formulation extends the traditional maximum entropy approach by associating uncertainty with the ME model. We therefore have bridged the maximum entropy and Bayesian approaches in a new context.

APPENDIX

Proof of Lemma 1. We show the result along the lines of Cover and Thomas (1991, pp. 228–229). Since the density is absolutely continuous, by the mean value theorem, there exists a point $\zeta_k^* \in (\zeta_{k-m}, \zeta_{k+m})$ such that

$$\Delta F_{m,k} = \int_{\zeta_{k-m}}^{\zeta_{k+m}} f(x) dx = \Delta \zeta_{m,k} f(\zeta_k^*).$$

In particular, $\Delta F_k = \Delta \zeta_k f(\zeta_k^*)$. Substituting in (16), we have

$$\begin{aligned} H_{m,q}(F) &= - \sum_{k=1}^q \Delta \zeta_k f(\zeta_k^*) \log \frac{\Delta \zeta_{m,k} f(\zeta_k^*)}{\Delta \zeta_{m,k}} \\ &= - \sum_{k=1}^q \Delta \zeta_k f(\zeta_k^*) \log f(\zeta_k^*) \\ &\rightarrow - \int f(x) \log f(x) dx \quad \text{as } \Delta \zeta_{m,k} \rightarrow 0 \quad \forall k. \end{aligned}$$

The limit is due to the fact that since the integral (3) exists, $f(x) \log f(x)$ is Riemann integrable.

The following results are needed for proof of Theorem 1.

Lemma 2. Let X be a random variable with Beta density

$$g(x) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} x^{u-1}(1-x)^{v-1}, \quad 0 \leq x \leq 1, \quad u > 0, \quad v > 0.$$

Then:

- (i) $E(\log X) = [\psi(u) - \psi(u+v)]$;
 (ii) $E(X \log X) = \frac{u}{u+v} [\psi(u+1) - \psi(u+v+1)]$;

Proof. Part (i) is well known. For (ii) note that for $\alpha = 1, 2, \dots$,

$$\begin{aligned} E[(X \log X)^\alpha] &= \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \int_0^1 (\log x)^\alpha x^{u+\alpha-1}(1-x)^{v-1} dx \\ &= \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \int_0^1 \frac{\partial^\alpha}{\partial u^\alpha} [x^{u+\alpha-1}(1-x)^{v-1}] dx \\ &= \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \frac{\partial^\alpha}{\partial u^\alpha} \int_0^1 x^{u+\alpha-1}(1-x)^{v-1} dx \\ &= \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \frac{\partial^\alpha}{\partial u^\alpha} \left[\frac{\Gamma(u+\alpha)\Gamma(v)}{\Gamma(u+v+\alpha)} \right]. \end{aligned}$$

The result is obtained upon differentiation and simplification and letting $\alpha = 1$.

Lemma 3. Let $X = (X_1, \dots, X_{q-1})'$ be a Dirichlet vector with density

$$g(x|u) = \frac{\Gamma\left(\sum_{k=1}^q u_k\right)}{\prod_{k=1}^q \Gamma(u_k)} \prod_{k=1}^{q-1} x_k^{u_k-1} \left\{ 1 - \sum_{k=1}^{q-1} x_k \right\}^{u_q-1}, \quad \sum_{k=1}^{q-1} x_k \leq 1, \quad x_k \geq 0, \\ u_k > 0, \quad k = 1, \dots, q.$$

Then

$$E_{g(\cdot|u)}[X_k \log T_k(X)] = \frac{u_k}{\sum_{k=1}^q u_k} E_{g(\cdot|v)}[\log T_k(X)],$$

where $T_k(X) = \sum_{j=k-a}^{k+b} X_j$ and the expectation in the right hand side is with respect to the Dirichlet density with parameters $v_j = u_k + \delta(j=k)$, δ being the indicator function.

Proof. Since $v_j = u_k + 1$ when $j = k$ and $v_j = u_k$ for all $j \neq k$, we have $\prod_{j=1}^{q-1} x_j^{v_j-1} = x_k \prod_{j=1}^{q-1} x_j^{u_j-1}$ and $\sum_{j=1}^q v_j = 1 + \sum_{k=1}^q u_k$. By recursion of the

gamma function we have,

$$\Gamma\left(\sum_{k=1}^q v_k\right) = \Gamma\left(1 + \sum_{k=1}^q u_k\right) = \left(\sum_{k=1}^q u_k\right) \Gamma\left(\sum_{k=1}^q u_k\right)$$

$$\prod_{k=1}^q \Gamma(v_k) = u_k \prod_{j=1}^q \Gamma(u_j).$$

Using the above relations in the expectation integral gives the result.

Proof of Theorem 1. (i) Write the quantized entropy as

$$H_{m,q}(F) = \sum_{k=1}^q \Delta F_k \log \Delta \xi_{m,k} - \sum_{k=1}^q \Delta F_k \log \Delta F_{m,k}.$$

The expected value of the first summation is obtained by noting that the increments $\Delta F_k, k = 1, \dots, q$ are distributed as Beta with parameters $\tilde{\mathcal{B}}\tilde{\Delta F}_k$ and $\tilde{\mathcal{B}}(1 - \tilde{\Delta F}_k)$. Thus $E(\Delta F_k) = \tilde{\Delta F}_k$.

The expected value of the second summation is obtained as follows. For $m = 0$, using part (ii) of Lemma 2 with $u = \tilde{\mathcal{B}}\tilde{\Delta F}_k$ and $v = \tilde{\mathcal{B}}(1 - \tilde{\Delta F}_k)$ gives the result. For $m \neq 0$, first use Lemma 3 with $\Delta F_{m,k} = \sum_{j=k-m+1}^{k+m} \Delta F_j = T_k(\Delta F)$. Then use the fact that the distribution of $T_k(X)$ is Beta with parameters $T_k(u)$ and $\sum_{k=1}^q u_k - T_k(u)$ and part (i) of Lemma 2. The result is obtained upon some simplifications.

(ii) Using the recursion $\psi(z + 1) = \psi(z) + z^{-1}$ (Abramowitz and Stegun, 1970) we can write (23) as

$$\tilde{H}_{m,q}(F) = \psi(\tilde{\mathcal{B}}) + \frac{1}{\tilde{\mathcal{B}}} - \sum_{k=1}^q \frac{\tilde{\Delta F}_k}{\tilde{\mathcal{B}}\tilde{\Delta F}_{m,k}} - \sum_{k=1}^q \tilde{\Delta F}_k [\psi(\tilde{\mathcal{B}}\tilde{\Delta F}_{m,k}) - \log \Delta \xi_{m,k}].$$

For large $\tilde{\mathcal{B}}$ the second and third terms vanish. Also for large $\tilde{\mathcal{B}}, \psi(\tilde{\mathcal{B}}) \approx \log \tilde{\mathcal{B}}$ (Abramowitz and Stegun, 1970), so

$$\tilde{H}_{m,q}(F) \approx \log \tilde{\mathcal{B}} - \sum_{k=1}^q \tilde{\Delta F}_k (\log \tilde{\mathcal{B}} + \log \tilde{\Delta F}_{m,k} - \log \xi_{m,k}),$$

which gives the result.

Proof of Theorem 2. For large n the sample dominates the prior in (22) and $\tilde{\Delta F}_{0,k} \rightarrow \widehat{\Delta F}_k$ as $n \rightarrow \infty$. Then the consistency of the histogram entropy (Hall and Morton, 1993) gives the result.

Proof of Theorem 3. For large n the sample dominates the prior in (22), i.e., $\Delta\tilde{F}_{m,i} \rightarrow \Delta\hat{F}_{m,i} = (2m)/n$ as $n \rightarrow \infty$. Since $n \rightarrow \infty$ and $m/n \rightarrow 0$, we can use the approximation (24) with $\Delta\tilde{F}_{m,i} = (2m)/n$ which gives $\tilde{H}_{m,n}(F) \xrightarrow{p} H_v(m, n)$. Then the result follows from the consistency of $H_v(m, n)$; see Vasicek (1976).

Proof of Remark 1. Since for $m = 0$, $\Delta F_{m,i} = \Delta F_i$ and n is large, the approximation (24) is applicable. Then the consistency of the empirical distribution implies

$$\tilde{H}_{m,n}(F) \xrightarrow{p} H_{m,n}(F), \quad \text{as } n \rightarrow \infty.$$

Lemma 1 gives the result.

Proof of Theorem 4. Let $\zeta_k = y_k, k = 1, \dots, q$. Then $H(\Delta F) = H(f_1, \dots, f_q)$. To see that the uniform prior guess is sufficient, let $n = 0$ and $\Delta F_k^* = q^{-1}$ in (26) and obtain the prior average entropy obtained by Campbell (1995), $E[H(\Delta F) | F_k^* = k/q, \mathcal{B}] = \psi(\mathcal{B} + 1) - \psi(\mathcal{B}/q + 1)$; i.e., $\rho = \mathcal{B}/q$. Also note that with the uniform best guess, $\alpha(y_k) = k\mathcal{B}/q$, thus $\alpha(y_k) - \alpha(y_{k-1}) = \mathcal{B}/q \equiv \rho$; i.e., (19) reduces to the symmetric Dirichlet prior (27). Conversely, the uniform distribution is implied by letting $\alpha(y_k) - \alpha(y_{k-1}) = \rho$ in (19) which implies that $\alpha(y_k) = k\rho$ and $\mathcal{B} = q\rho$.

Proofs of Theorem 5. As $n \rightarrow \infty$ the quantized moments approach to the sample moments. The consistency of the sample moments implies the consistency of $H(F^* | \theta_q)$. Then results follow from the uniqueness of the ME density and Theorems 2 and 3.

Proof of Remark 2. Use the same argument as proof Theorem 5.

ACKNOWLEDGMENTS

We acknowledge with thanks the comments received from Amos Golan, and two referees on a previous draft of this article. Their comments enhanced our understanding of the topic and were very helpful in improving the presentations. We are, however, responsible for any remaining shortcomings.

REFERENCES

- Abramowitz, M., Stegun, I. A. (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19:716–723.

- Arizono, I., Ohta, H. (1989). A test for normality based on Kullback–Leibler information. *The American Statistician* 34:20–23.
- Beirlant, J., Dudewicz, E. J., Györfi, L., van der Meulen, E. C. (1997). Non-parametric entropy estimation: an overview. *International Journal of Mathematical and Statistical Sciences* 6:17–39.
- Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics* 7:686–690.
- Bowyer, A. (1981). Computing Dirichlet tessellations. *The Computer Journal* 24:162–166.
- Campbell, L. L. (1995). Averaging entropy. *IEEE Transactions on Information Theory IT* 41:338–339.
- Carota, C., Parmigiani, G., Polson, N. G. (1996). Diagnostic measures for model criticism. *Journal of the American Statistical Association* 91:753–762.
- Choi, B., Kim, K. (2006). Testing goodness-of-fit for Laplace distribution based on maximum entropy. *Statistics* 40:517–531.
- Clarke, B. (1996). Implications of reference priors for prior information and for sample size. *Journal of the American Statistical Association* 91:173–184.
- Clarke, B., Gustafson, P. (1998). On the sensitivity of posterior distribution to its inputs. *Journal of Statistical Planning and Inference* 71:137–150.
- Cover, T. M., Thomas, J. A. (1991). *Elements of Information Theory*. New York: John Wiley.
- Dadpay, A., Soofi, E. S., Soyer, R. (2007). Information measures for generalized gamma family. *Journal of Econometrics* 138:568–585.
- Dempster, A. P. (1974). The direct use of likelihood for significance testing. In: Proceedings of Conference on Foundational Questions in Statistical Inference. pp. 335–352.
- DeWaal, D. J. (1996). Goodness of fit of the generalized extreme value distribution based on the Kullback–Leibler information. *South African Statistical Journal* 30:139–153.
- Dudewicz, E. J., Van Der Meulen, E. C. (1981). Entropy-based tests of uniformity. *Journal of the American Statistical Association* 76:967–974.
- Dudewicz, E. J., Van Der Meulen, E. C. (1987). Empiric entropy, a new approach to nonparametric entropy estimation. In: *New Perspectives in Theoretical and Applied Statistics*. New York: Wiley, pp. 207–227.
- Ebrahimi, N. (1997). Testing whether lifetime distribution is decreasing uncertainty. *Journal of Statistical Planning and Inference* 64:9–19.
- Ebrahimi, N. (1998). Testing for exponentiality of the residual lifetime based on dynamic Kullback–Leibler information. *IEEE Transactions on Reliability* R-47:197–201.
- Ebrahimi, N. (2001). Testing for uniformity of the residual lifetime based on dynamic Kullback–Leibler information. *Annals of Institute of Statistical Mathematics* 53:325–337.
- Ebrahimi, N., Habibullah, M., Soofi, E. S. (1992). Testing exponentiality based on Kullback–Leibler information. *Journal of Royal Statistical Society B* 54:739–748.
- Ebrahimi, N., Pflughoeft, K., Soofi, E. S. (1994). Two measures of sample entropy. *Statistics and Probability Letters* 20:225–234.
- Ebrahimi, N., Maasoumi, E., Soofi, E. S. (1999). Ordering univariate distributions by entropy and variance. *Journal of Econometrics* 90:317–336.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 2:209–230.
- Gill, C. A., Joanes, D. N. (1979). Bayesian estimation of Shannon’s index of diversity. *Biometrika* 66:81–85.
- Gokhale, D. V. (1983). On entropy-based goodness-of-fit tests. *Computational Statistics and Data Analysis* 1:157–165.
- Hall, P., Morton, S. C. (1993). On the estimation of entropy. *Annals of Institute of Mathematical Statistics* 45:69–88.
- Inverardi, P. L. N. (2003). MSE comparison of some different estimators of entropy. *Communications in Statistics – Simulation and Computation* 32:17–30.
- Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Annals of Institute of Mathematical Statistics* 41:683–697.
- Kraskov, A., Stögbauer, H., Grassberger, P. (2004). Estimating mutual information. *Physical Review E* 69: Art. No. 066138.
- Mazzuchi, T. A., Soofi, E. S., Soyer, R. (2000). Computations of maximum entropy Dirichlet for modeling lifetime data. *Computational Statistics and Data Analysis* 32:361–378.
- Mudholkar, G. S., Tian, L. (2002). An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test. *Journal of Statistical Planning and Inference* 102:211–221.

- Park, S. G. (1999). A goodness-of-fit test for normality based on the sample entropy of order statistics. *Statistics and Probability Letters* 44:359–363.
- Park, S. G. (2005). Testing exponentiality based on the Kullback–Leibler information with the type II censored data. *IEEE Transactions on Reliability* 54:22–26.
- Park, S., Park, D. (2003). Correcting moments for goodness of fit tests based on two entropy estimates. *Journal of Statistical Computation and Simulation* 73:685–694.
- Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science* 19:652–662.
- Soofi, E. S. (1997). Information theoretic regression methods. In: *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems 12*. Greenwich, CT: JAI Press, pp. 25–83.
- Soofi, E. S., Retzer, J. J. (2002). Information indices: unification and applications. *Journal of Econometrics* 107:17–40.
- Soofi, E. S., Ebrahimi, N., Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association* 90:657–668.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussions). *Journal of the Royal Statistical Society, Series B* 64:583–616.
- Taufer, E. (2002). On entropy based tests for exponentiality. *Communications in Statistics-Simulation and Computation* 31:189–200.
- Theil, H. (1980). The entropy of the maximum entropy distribution. *Economics Letters* 5:145–148.
- Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of Royal Statistical Society B* 38:54–59.
- Yuan, A., Clarke, B. (1999). An information criterion for likelihood selection. *IEEE Transactions on Information Theory IT* 45:562–571.
- Zellner, A. (1991). Bayesian methods and entropy in economics and econometrics. In: *Maximum Entropy and Bayesian Methods*. Netherlands: Kluwer, pp. 17–31.
- Zellner, A. (1996). Models, prior information, and Bayesian analysis. *Journal of Econometrics* 75:51–68.