

Measuring Informativeness of Data by Entropy and Variance *

Nader Ebrahimi

Division of Statistics

Northern Illinois University

DeKalb, IL 60115

nader@math.niu.edu

Esfandiar Maasoumi

Department of Economics

Southern Methodist University

Dallas, TX 75275-0496

maasoumi@post.cis.smu.edu

Ehsan S. Soofi

School of Business Administration

University of Wisconsin-Milwaukee

P.O. Box 742, Milwaukee, WI 53201

esoofi@csd.uwm.edu

DATE OF DRAFT: April 20, 1998

*This paper is a contribution in honor of Professor Camilo Dagum for his many deep insights and contributions to economics.

Abstract

Entropy and Variance are “indices” that have been used to measure dispersion, volatility, risk, uncertainty, and information. Variance has been prominent but the use of entropy is growing rapidly. This paper examines the use of variance and entropy for measuring informativeness of data in the Bayesian setting. Our main findings may be summarized as follows. For binary data, when the prior is uniform, all data are “informative”. But we find that with other conjugate priors, such as Jeffreys’ prior, “surprises” are possible since the posterior entropy may be increased while variance is always reduced. With Zellner’s Maximal Data Information Prior (Zellner 1971), however, all binary sample are also informative. In our second case of exponential data, variance and entropy may differ in their verdict for finite sample sizes. For large samples, however, both will agree that the data are informative. In our last case of Gaussian data and conjugate priors, when the variance is known, the data are informative on the mean, variance and entropy agree on their verdicts. But when the mean is known, variance and entropy may not agree on the informativeness of the data on the unknown variance is finite samples. Predictably, they do agree in large samples.

Key words: Risk; volatility; entropy; information theory; ordering relations.

1 Introduction

Measuring informativeness of data/news is particularly important as it quantifies the amount of “learning”. This is central to scientific progress as well as to assessing the direction and value of “information” and technologies. As is the case with all “indices”, the desirability of any measure of information depends on at least two considerations: First is the inference/investigative technique that would utilize the information. The second is the distributional characteristics of the information which is to be summarized. As examples, least squares techniques are, by design, incapable of utilizing any information other than the “variation” in a distribution/data. And, the Gaussian distributions/data are entirely characterized by the first two moments; any index will thus be a function of the same moments. These two considerations need to be borne in mind when contrasting entropy and variance as indices of informativeness or uncertainty.

Bayesian learning and updating is one of the most elegant and prevalent of the formalisms in all of science. It provides for a particularly efficient means of assessing the value of new information. It is in this context that we measure informativeness. A Bayesian analysis includes a likelihood function, $f(x|\theta)$ and a prior distribution $P(\theta)$ which maps the analyst’s uncertainty about the parameter θ . The prior is updated by the observed data $x = (x_1, \dots, x_n)$ into a posterior distribution $P(\theta|x)$. The comparison of uncertainty before and after observing the data is of interest.

We are concerned with the problem of evaluating the informativeness of a given data set. Our results pertain to the post data stage; Abel and Singpurwalla (1994) provide an interesting example of the problem in a reliability analysis context. The problem of measuring the informativeness of a given data set at the post data stage is different than the more extensively analyzed pre-observation, or the design stage in which the “expected information” relative to the variable space is measured. It is well-known that, on average, samples are informative according to both variance and entropy (see, e.g., Lindley 1956). However, this may not be the case for some data configurations and prior distributions.

The outline of this paper is as follows. In section 2, we provide an overview of some information functions for quantifying the information content of a given data and the ex-

pected information in the data. Section 3 discusses the informativeness of binary data under two types of non-informative priors for the Bernoulli parameter. Section 4 discusses the informativeness of exponential data under the conjugate prior. Section 5 discusses the informativeness of Gaussian data under the conjugate priors for the mean and variance. Section 6 concludes.

2 The role of information theory

The literature in economics, econometrics, and statistics has witnessed a great rise in the use of information theory concepts and measures in the last decade or so. The axiomatic appeal and the role played by entropy as a criterion function in deriving optimal measures and Maximum Entropy (ME) distributions, explain the recent abundance of entropy-based methods in econometrics and other areas; see, e.g., Soofi (1990, 1994, 1997), Zellner (1988, 1991, 1996a, 1996b, 1997), Maasoumi (1993, 1997), Ryu (1993), Golan, Judge, and Miller (1996), Fomby and Hill (1997), and references therein. Holm (1993) has recently developed ME Lorenz curves. Stutzer (1995, 1996) has introduced information theoretic financial indices. In the Bayesian Method of Moments (BMOM) approach, for example, the ME is the vehicle for generating post-data distributions for the *structural parameters* of econometric models and for prediction; see Zellner (1994, 1996a-b, 1997), Zellner and Sacks (1996), Zellner, Tobias, and Ryu (1997), Tobias and Zellner (1997).

Entropy of a parameter θ with an absolutely continuous prior probability distribution $P(\theta)$ over the parameter space Θ is defined by

$$H(\Theta) \equiv H[p(\theta)] = - \int_{\Theta} p(\theta) \log p(\theta) d(\theta), \quad (1)$$

where $p(\theta)$ is the probability density function of P .

The entropy measures the “uniformity” of a distribution. $H(\Theta)$ increases as $p(\theta)$ approaches a uniform distribution. Consequently, the concentration of probabilities decreases and it becomes more difficult to predict θ . In this sense $H(\Theta)$ is a measure of *uncertainty* associated with $p(\theta)$. The negative entropy $-H(\Theta)$ is used as a measure of information (Zellner 1971); see Soofi and Gokhale (1997) for a justification.

A conditional entropy is obtained by using a conditional density in (1). The posterior entropy is given by the conditional entropy $H(\Theta|\mathbf{x}) = H[p(\theta|\mathbf{x})]$ and the entropy of the sampling distribution is $H(X) = H[f(\mathbf{x}|\theta)]$.

A conditioning may increase or decrease the entropy. The expected conditional entropy with respect to the sampling distribution is $E_{\mathbf{x}}[H(\Theta|\mathbf{x})] \leq H(\Theta)$; the equality holds if and only if Θ and \mathbf{X} are stochastically independent. That is, on average, conditioning decreases the entropy as is the case for variance.

In Bayesian statistics, the information about a parameter is quantified by a discrepancy measure between the posterior and prior distributions; see, e.g., Lindley (1956), Goel and DeGroot (1979), Zellner (1971, 1984), and Goel (1983). We measure the information provided by the actual data \mathbf{x} about a parameter $\theta \in \Theta$ by the entropy difference

$$\vartheta(\Theta|\mathbf{x}) = H(\Theta) - H(\Theta|\mathbf{x}).$$

The informativeness of the data is indicated by the sign of $\vartheta(\Theta|\mathbf{x})$. When $\vartheta(\Theta|\mathbf{x}) > 0$, the uncertainty is reduced by the data and the sample is said to be informative; otherwise the data is said to have produced a “surprise” (Lindley 1956).

Following Lindley (1956), the information function that has been widely used for comparison of experiments at the planning stage for the purpose of data collection is the mutual information defined as:

$$\begin{aligned} \vartheta(\Theta \wedge \mathbf{X}) &\equiv E_{\mathbf{x}}[\vartheta(\Theta|\mathbf{x})] \\ &= H(\Theta) - E_{\mathbf{x}}[H(\Theta|\mathbf{x})] \end{aligned}$$

See Soofi (1997) for details and applications.

The mutual information may be written in terms of the following Kullback-Leibler discrimination information functions:

$$\begin{aligned} \vartheta(\Theta \wedge \mathbf{X}) &= K[f(\mathbf{x}, \theta) : f(\mathbf{x})p(\theta)] \\ &= \int_{\mathbb{R}^n} \int_{\Theta} f(\mathbf{x}, \theta) \log \frac{f(\mathbf{x}, \theta)}{f(\mathbf{x})p(\theta)} d\theta d\mathbf{x} \\ &= E_{\mathbf{x}}\{K[p(\theta|\mathbf{x}) : p(\theta)]\} \end{aligned}$$

The mutual information $\vartheta(\Theta \wedge \mathbf{X})$ provides a measure of expected information discrepancy between the posterior and prior distributions, which is the expected information in yet

unobserved data \mathbf{X} about the parameter. Note that $\vartheta(\Theta \wedge \mathbf{X}) \geq 0$, with equality if and only if $f(\theta, \mathbf{x}) = p(\theta)f(\mathbf{x})$. Accordingly, $\vartheta(\Theta \wedge \mathbf{X})$ is also a measure of stochastic dependency between the two variables.

In the traditional statistics, the variance is used for measuring uncertainty. The widespread use of variance for measuring uncertainty is rooted in statistical estimation (Fisher 1921). In statistical estimation, Fisher's information is defined as

$$\mathcal{F}(\theta) \equiv \mathcal{F}[f(x|\theta)] = -E_{x|\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right].$$

$\mathcal{F}(\theta)$ is a measure of information in X , i.e., in $f(x|\theta)$ about the parameter θ , in the sense that $\mathcal{F}(\theta)$ quantifies "the ease with which a parameter can be estimated" by x (Lehmann 1983, p. 120). Inherent in this interpretation are the facts that: (a) X is an unbiased and efficient estimator of θ , so $V(X|\theta) = [\mathcal{F}(\theta)]^{-1}$, and (b) under $f(x|\theta)$, the probabilities are concentrated around the mean value θ .

From the information-theoretic viewpoint, the Fisher information \mathcal{F} is a second order approximation to the discrimination information function $K[f(x|\theta) : f(x|\theta + \Delta\theta)]$ where θ and $\theta + \Delta\theta$ are two neighboring points in the parameter space and the two distributions f_θ and $f(x|\theta), f(x|\theta + \Delta\theta)$ are two densities in the same parametric family (Kullback 1959). Lindley (1961) showed that ignorance between two neighboring values θ and $\Delta\theta$ in the parameter space implies that $\vartheta(\Theta \wedge X) \approx 2(\Delta\theta)^2 \mathcal{F}(\theta)$.

Ebrahimi, Maasoumi, and Soofi (1998) explore the relationship in terms of an approximation of the density, and discuss the equivalence of entropy and variance orderings implied by a more general partial order relation between random variables. They also identify a few transformations of continuous random variables that preserve the equivalence of variance and entropy orderings, and offer some results on the equivalence of entropy and variance orderings for well-known families of continuous and discrete distributions.

Zellner (1971) defined an information function for quantifying the information in the data x about a parameter θ with the prior $p(\theta)$, which may be written as:

$$\begin{aligned} G[p(\theta)] &= E_\theta \{ H[p(\theta)] - H[f(x|\theta)] \} \\ &= E_\theta \{ K[f(x|\theta) : p(\theta)] \} \\ &= \vartheta(\Theta \wedge X) + H(\Theta) - H(X). \end{aligned} \tag{2}$$

Zellner proposed $G[p(\theta)]$ as a criterion function for developing prior distributions that are maximally committed to the data. The prior $p^*(\theta)$ that maximizes $G[p(\theta)]$ is referred to as the *Maximal Data Information Prior (MDIP)*. The first equation in (2) is the *a priori* expected information in the data-generating density (likelihood function) which is “purified” from the information in the prior. The second equation in (2) shows that $G[p(\theta)]$ is the *a priori* expected information for discrimination between the data-generating distribution and the prior. The MDIP gives explicit solutions in many problems and is capable of including side information in terms of moment constraints on $p(\theta)$; see Zellner (1991) for details.

We use the following notations in the sequel. Let F_1 and F_2 be two distributions with entropies H_1, H_2 and variances V_1, V_2 . Then, the *Variance Ordering* $V_1 \leq V_2$ will be denoted by $P_1 \overset{V}{<} P_2$ and the *Entropy Ordering* $H_1 \leq H_2$ will be denoted as $P_1 \overset{E}{<} P_2$. When variance and entropy order the two distributions similarly, we write $P_1 \overset{EV}{<} P_2$.

3 Informativeness of Binary Data

The likelihood function for the Bernoulli parameter based on the binary data, x_1, \dots, x_n , is

$$p(y|\theta) = \theta^y(1 - \theta)^{n-y}, \quad y = \sum_{i=1}^n x_i, \quad x_i = 0, 1, \quad 0 \leq \theta \leq 1. \quad (3)$$

We wish to evaluate the informativeness of the data about the parameter θ . We consider two classes of prior distributions for θ .

3.1 Conjugate Priors

The conjugate family of priors for the likelihood function (3) is $P(\theta) = \text{Beta}(a, b)$. The posterior distribution is $P(\theta|n, y) = \text{Beta}(y + a, n - y + b)$. Two important examples of conjugate priors are the uniform prior $P(\theta) = \text{Beta}(1, 1)$ and Jeffreys' invariant prior $P(\theta) = \text{Beta}(.5, .5)$.

Under the uniform prior, the posterior distribution is $P(\theta|n, y) = \text{Beta}(y + 1, n - y + 1)$. It is well known that among all distributions with a given support, the uniform distribution has the maximum entropy. Thus under the uniform prior, any sample is informative due to

reductions in posterior entropy and variance.

Under Jeffreys' prior, the posterior distribution is $P(\theta|n, y) = \text{Beta}(y + .5, n - y + .5)$. In this case, however, not all samples are informative about θ if uncertainty is measured by entropy.

Lindley (1957) showed that for the Beta family $\text{Beta}(\alpha, \beta)$, when α and β are large, entropy and Fisher information (variance) behave similarly. Ebrahimi, Maasoumi, and Soofi (1998) showed that $P_1 \stackrel{EV}{<} P_2$ holds for $\alpha_1 < \alpha_2$ and $\beta_1 < \beta_2$ when $(\alpha, \beta) \in S_\alpha \cap S_\beta$, each region defined as follows:

$$S_\alpha = \left\{ (\alpha, \beta) : \alpha > 1 - \frac{(\beta - 1)\psi_\alpha(\alpha + \beta)}{\psi_\alpha(\alpha) - \psi_\alpha(\alpha + \beta)} \right\}$$

$$S_\beta = \left\{ (\alpha, \beta) : \beta > 1 - \frac{(\alpha - 1)\psi_\beta(\alpha + \beta)}{\psi_\beta(\beta) - \psi_\beta(\alpha + \beta)} \right\},$$
(4)

where $\psi_z(z)$ is the derivative of the digamma function $\psi(z)$.

When n and y are large, (4) holds. In some small samples, however, variance and entropy may give opposite assessments of informativeness of data under the Jeffreys' prior.

Table 1 shows $H(\theta) - H(\theta|n, y)$ and $V(\theta) - V(\theta|n, y)$ for $n \leq 5$. We note that $V(\theta) - V(\theta|n, y) > 0$ for all y , an indication of monotone decrease in spread of the posterior distribution around the mean. But for $n \leq 4$, $P(\theta) \stackrel{E}{<} P(\theta|n, y)$ holds only when y is near 0 or

Table 1. Posterior Entropy and Variance of Binomial Experiments
Based on Jeffreys' Prior, and y Successes in n Trials.

y	$H(\theta) - H(\theta n, y)$					$V(\theta) - V(\theta n, y)$				
	n					n				
	1	2	3	4	5	1	2	3	4	5
0	0.306	0.708	1.004	1.234	1.423	0.062	0.090	0.103	0.110	0.114
1	0.306	-0.194	-0.053	0.114	0.267	0.062	0.062	0.078	0.090	0.098
2		0.708	-0.053	-0.042	0.048		0.090	0.078	0.083	0.090
3			1.004	0.114	0.048			0.103	0.090	0.090
4				1.234	0.267				0.110	0.098
5					1.423					0.114

near n . For $y = n/2$, $n = 2, 4$ and for $y = 1, 2$, $n = 3$, $H(\theta) - H(\theta|n, y) < 0$. Thus under the conjugate family of priors, binary data can produce a “surprise” according to entropy, but not according to variance.

3.2 Maximal Data Information Prior (MDIP)

The MDIP for the Bernoulli parameter is

$$p^*(\theta) = C_0 \theta^\theta (1 - \theta)^{1-\theta}, \quad 0 \leq \theta \leq 1. \quad (5)$$

where the normalizing constant $C_0 \approx 1.6185$ found by a numerical evaluation; see Zellner (1984).

The density (5) is symmetric around $\theta = 0.5$ which is the minimum. Moreover, $p(0.5) = 0.5C_0$, and $\lim_{\theta \rightarrow 0} p^*(\theta) = \lim_{\theta \rightarrow 1} p^*(\theta) = C_0$. Thus, the MDIP gives twice as much probabilities to the values of θ near each end point, zero and one, as to the central values. However, the MDIP is not as extreme as the Jeffreys' prior in assigning high probabilities to the end values.

The least informative data configuration is when $n = 2k$ and $y = k$, $k = 0, 1, 2, \dots$. In this case, the posterior density is

$$p(\theta|n = 2k, y = k) = C_k \theta^{k+\theta} (1 - \theta)^{k+1-\theta}, \quad (6)$$

where C_k must be found numerically by evaluating

$$C_k^{-1} = \int_0^1 \theta^{k+\theta} (1 - \theta)^{k+1-\theta} d\theta.$$

Note that C_k^{-1} , $k = 0, 1, 2, \dots$ is a decreasing sequence, so C_k , $k = 0, 1, 2, \dots$ is an increasing sequence. For $k > 0$, $\lim_{\theta \rightarrow 0} p(\theta|n = 2k, y = k) = \lim_{\theta \rightarrow 1} p(\theta|n = 2k, y = k) = 0$.

Entropy of (6) is given by

$$H(\theta|n = 2k, y = k) = -(\log C_k + 2C_k D_k),$$

where

$$D_k = \int_0^1 (k + \theta) \log(\theta) \theta^{k+\theta} (1 - \theta)^{k+1-\theta} d\theta.$$

Although, D_k , $k = 0, 1, 2, \dots$ is a decreasing sequence, it can be shown that the product $C_k D_k$, $k = 0, 1, 2, \dots$ is an increasing sequence. Hence, $H(\theta|n = 2k, y = k)$, $k = 0, 1, 2, \dots$ is a *decreasing* sequence.

The density (6) is symmetric around $\theta = 0.5$. Variance is given by

$$V(\theta|n = 2k, y = k) = C_k \int_0^1 \theta^{k+\theta+2} (1-\theta)^{k+1-\theta} d\theta - \frac{1}{4}.$$

It can be shown that $V(\theta|n = 2k, y = k)$, $k = 0, 1, 2, \dots$ is also a *decreasing* sequence.

Figure 1 shows Jeffreys' prior $Beta(.5, .5)$ (dash-2 points), Zellner's MDIP (dash-1 point), and the associated posterior densities for $n = 2k$, $k = 1, 2, 3$. The solid curves are the posteriors based on the MDIP. All three have less entropy and variance as compared with the prior. The two dashed curves and the dash-3 dots curve are the Beta posteriors based on Jeffreys' prior. The first two have larger entropies and the third one has smaller entropy than $Beta(.5, .5)$. Since under $Beta(.5, .5)$ the probability is heavily concentrated at the tails of the distribution, the prior has a larger variance than all posteriors. For $n = 2, 4$, the probability is less concentrated under the posteriors than the prior.

We conclude that under Zellner's MDIP (5), all binary samples are informative whether measure uncertainty by entropy or by variance; i.e., under the MDIP, no sample may produce a "surprise".

4 Informativeness of Exponential Data

For exponentially distributed occurrence times with rate θ , the conjugate prior is Gamma, $P(\theta) = G(\alpha, \beta)$. Note that θ is the precision parameter of the exponential distribution, whereas β is the scale parameter of the Gamma prior for θ . The posterior distribution is $P(\theta|\mathbf{x}) = G[n + \alpha, \beta(1 + \beta T_n)^{-1}]$, where $T_n = \sum x_i$. That is, the sample increases the shape parameter and decreases the scale parameter of the prior gamma distribution.

In order to determine the informativeness of the sample, we note that $P(\theta|\mathbf{x}) \stackrel{V}{<} P(\theta)$ if and only if

$$\log(1 + \beta T_n) \geq (1/2)\log(1 + n/\alpha), \quad (7)$$

and that $P(\theta|\mathbf{x}) \stackrel{E}{<} P(\theta)$ if and only if

$$\log(1 + \beta T_n) \geq \log[\Gamma(\alpha + n)/\Gamma(\alpha)] + (1 - \alpha)[\psi(\alpha + n) - \psi(\alpha)] - n\psi(\alpha + n) + n. \quad (8)$$

Thus, $P(\theta|\mathbf{x}) \stackrel{EV}{<} P(\theta)$ does not always hold. However, we can establish $P(\theta|\mathbf{x}) \stackrel{EV}{<} P(\theta)$ for large samples. Using the asymptotic approximations,

$$\log[\Gamma(z)] \approx 1/2\log(2\pi) - z + (z - 1/2)\log(z) \quad (9)$$

and

$$\psi(z) \approx \log(z) - 1/2z^{-1}, \quad (10)$$

the condition (8) reduces to

$$\log(1 + \beta T_n) \geq (1/2)\log(1 + n/\alpha) + (1 - \alpha)/(2\alpha) - (\alpha - 1/2)\log \alpha + o(n). \quad (11)$$

As $n \rightarrow \infty, T_n/n \rightarrow \theta^{-1}$, and $E_\theta(\theta^{-1}) = (\alpha - 1)\beta^{-1}$, where the expectation is taken with respect to the prior distribution $P(\theta)$. We find that for large n and $\alpha > 1$, $\log(1 + \beta T_n) \rightarrow \log[1 + (\alpha - 1)n]$. Thus, averaged over the parameter space, the conditions (7) and (11) are asymptotically satisfied.

We remark that the notion of average entropy has been used before. The prior and posterior information in the density $p(x|\theta)$ are defined by the prior and posterior average (negative) entropies $-E_{P(\theta)}[H(X|\theta)]$ and $-E_{P(\theta|x)}[H(X|\theta)]$; see Zellner (1971, 1991) and Zellner (1988, 1991). The notion of average entropy is also used in the entropy estimation context by Gill and Joanes (1979), Mazzuchi, Soofi, and Soyer (1997), and in the information theory literature by Campbell (1995).

5 Informativeness of Gaussian Data

For samples from the Gaussian distribution, $f(x|\theta, \sigma^2)$ with known σ^2 , the conjugate family of priors for θ is $P(\theta) = N(\alpha, \beta^2)$. The posterior distribution is $P(\theta|\mathbf{x}) = N[(\sigma^2\alpha + T_n\beta^2)/(\sigma^2 + n\beta^2), \beta^2\sigma^2/(\sigma^2 + n\beta^2)]$. Clearly, $V(\theta) > V(\theta|x)$ and $H(\theta) > H(\theta|x)$ for all samples x . Thus all samples are informative about the normal mean according to both variance and entropy. This seems to be more of an exception than a rule.

For $f(x|\mu, \theta) = N(\mu, \theta)$ with known μ , the conjugate family is the Inverse Gamma, $P(\theta) = IG(\alpha, \beta)$ and the posterior is $P(\theta|\mathbf{x}) = IG(\alpha + n/2, \beta + Q_n)$, where $Q_n = 1/2 \sum (x_i - \mu)^2$. In this case, we also note that $P(\theta|\mathbf{x}) \stackrel{V}{<} P(\theta)$, if and only if

$$\log(1 + Q_n/\beta) < \log[(\alpha + n/2 - 1)/(\alpha - 1)] + 1/2 \log[(\alpha + n/2 - 2)/(\alpha - 2)], \quad (12)$$

and $P(\theta|\mathbf{x}) \stackrel{E}{<} P(\theta)$, if and only if

$$\log(1 + Q_n/\beta) < \log[\Gamma(\alpha)/\Gamma(\alpha + n/2)] + (\alpha + 1)[\psi(\alpha + n/2) - \psi(\alpha)] + (n/2)[\psi(\alpha + n/2) - 1]. \quad (13)$$

Here, we also note that $P(\theta|\mathbf{x}) \stackrel{EV}{<} P(\theta)$ does not always hold.

• For large n , the condition (12) reduces to

$$\log(1 + Q_n/\beta) < (3/2) \log(\alpha + n/2) - [\log(\alpha - 1) + 1/2 \log(\alpha - 2)]. \quad (14)$$

Using the asymptotic approximations (9) and (10), the condition (13) reduces to

$$\log(1 + Q_n/\beta) < (3/2) \log(\alpha + n/2) + \log[\Gamma(\alpha)/\sqrt{2\pi}] - (\alpha + 1)\psi(\alpha) + \alpha - 1/2 + o(n). \quad (15)$$

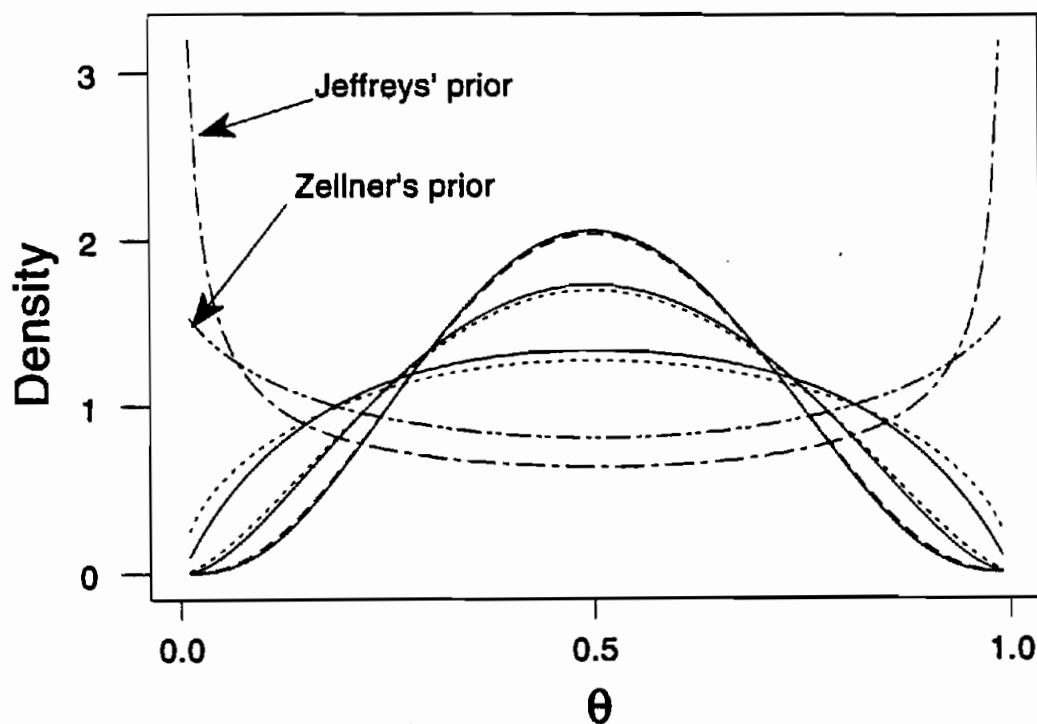
As $n \rightarrow \infty$, $Q_n/n \rightarrow \theta$, and $E_\theta(\theta) = (\alpha - 1)\beta$, where the expectation is taken with respect to the prior distribution $P(\theta)$. We find that for large n and $\alpha > 1$, $\log(1 + Q_n/\beta) \rightarrow \log[1 + (\alpha - 1)n]$. Thus, averaged over the parameter space, the conditions (14) and (15) are asymptotically satisfied.

6 Conclusions

An important example was given of Bayesian prior distributions specifying situations in which variance and entropy may or may not agree. For binary data, under a conjugate prior, the posterior entropy may increase or decrease as compared with the prior entropy, but the posterior variance always decreases. With Zellner's Maximal Data Information Prior (MDIP), both the posterior entropy and variance decrease for all binary data. Entropy analysis shows that while the MDIP shares a feature of Jeffreys' prior which assigns relatively higher probabilities to the end points than to the middle, it is not as extreme as Jeffreys' prior under which some samples ought to be evaluated as reducing information about the Bernoulli parameter!

Entropy and variance may also give opposite assessment of informativeness of exponential data. For the Gaussian case with a *known* scale parameter, under the conjugate prior for the mean, the entropy and variance concur that every data set is informative about the mean. But for the case of unknown scale and unknown mean, the two measures may give opposite assessment.

Figure 1. Jeffreys' and Zellner's Priors for the Bernoulli Parameter and Corresponding Posterior Distributions for $n = 2k$, $k = 1, 2, 3$.



- Dash-dot curve is the Jeffreys' invariant prior $Beta(.5, .5)$ for the Bernoulli parameter θ .
 - Dot curves are $Beta(1.5, 1.5)$ and $Beta(2.5, 2.5)$ posterior densities for $n = 2k$, $y = k = 1, 2$, which have higher entropies and lower variances than the Jeffreys' prior.
 - Dash curve is $Beta(3.5, 3.5)$ posterior density for $n = 6$, $y = 3$, which has a lower entropy and a lower variance than the Jeffreys' prior.

- Dash-2 dots curve is the Maximal Data Information Prior for the Bernoulli parameter θ .
 - Solid curves are posterior densities for $n = 2k$, $y = k = 1, 2, 3$, which have lower entropies and lower variances than the Maximal Data Info Prior.

References

- Abel, P. S. and N.D. Singpurwalla (1994) "To Survive or to Fail: That is the Question", *The American Statistician*, 48, 18-21.
- Ebrahimi, Maasoumi, and Soofi (1998) "Ordering Univariate Distributions by Entropy and Variance", *Journal of Econometrics*, to appear.
- Fisher, R. A. (1921) "On Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society of London, Ser. A*, 222, 309-368.
- Fomby, T. B. and R. C. Hill (1997) *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, Vol. 12, Greenwich CT: JAI Press.
- Goel, P.K. (1983) "Information Measures and Bayesian Hierarchical Models", *Journal of the American Statistical Association*, 78, 408-410.
- Goel, P.K. and M. H. DeGroot (1979) "Comparison of Experiments and Information Measures", *The Annals of Statistics*, 7, 1066-1077.
- Gill, C.A. and Joanes, D.N. (1979) "Bayesian Estimation of Shannon's Index of Diversity", *Biometrika*, 66, 81-85.
- Golan, A., Judge, G., and D. Miller (1996) *Maximum Entropy Econometrics*, New York: Wiley.
- Holm J. (1993) "Maximum Entropy Lorenz Curves", *Journal of Econometrics*, 59, 377-389.
- Kullback, S. (1959) *Information Theory and Statistics*, N.Y.: Wiley (reprinted in 1968 by Dover).
- Lehmann, E. L. (1983) *Theory of Point Estimation*, N.Y.: Wiley.
- Lindley, D.V. (1956) "On a Measure of Information Provided by an Experiment", *The Annals of Mathematical Statistics*, 27, 986-1005.
- Lindley, D.V. (1957) "Binomial Sampling Schemes and the Concept of Information", *Biometrika*, 44, 179-186.

- Lindley, D. V. (1961) "The Use of Prior Probability Distributions in Statistical Inference and Decision", *Proceedings of the Fourth Berkeley Symposium*, 1, 436-468, Berkeley: UC Press.
- Maasoumi, E. (1993) "A Compendium to Information Theory in Economics and Econometrics", *Econometric Reviews*, 12(2), 137-181.
- Maasoumi, E. (1997) "Empirical Analyses of Inequality and Welfare," in M.H. Pesaran and P. Schmidt (eds), *Handbook of Applied Microeconometrics*, Basil Blackwell.
- Mazzuchi, T. A., Soofi, E.S., and R. Soyer (1997) "Bayes Estimate of Entropy", under review.
- Ryu, H. K. (1993) "Maximum Entropy Estimation of Destiny and Regression Functions", *Journal of Econometrics*, 56, 397-440.
- Shannon, C. E. (1948) "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27, 379-423.
- Soofi, E. S. (1990) "Effects of Collinearity on Information About Regression Coefficients", *Journal of Econometrics*, 43, 255-274.
- Soofi, E. S. (1994) "Capturing the intangible concept of Information", *Journal of the American Statistical Association*, 89, 1243-1254.
- Soofi, E. S. (1997) "Information Theoretic Regression Methods", in *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, 12, T. B. Fomby and R. C. Hill (eds.), 25-83, Greenwich, CT: JAI Press.
- Soofi, E. S. and Gokhale, D.V. (1997) "Information Theoretic Methods for Categorical Data", in *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, 12, T. B. Fomby and R. C. Hill (eds.), 107-134, Greenwich, CT: JAI Press.
- Stutzer, M. (1995) "A Bayesian Approach to Diagnostics of Asset Pricing Models", *Journal of Econometrics*, 68, 367-397.

- Stutzer, M. (1996) "An Information-Theoretic Index of Risk in Financial Markets", in *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, D.A. Berry, K.M. Chanloner, and J.K. Geweke (eds.), New York: Wiley.
- Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley (reprinted in 1996 by Wiley)
- Zellner, A. (1984) *Basic Issues in Econometrics*, Chicago: University of Chicago Press.
- Zellner, A. (1988) "Optimal Information Processing and Bayes' Theorem" (with discussion), *The American Statistician*, 42, 278-284.
- Zellner, A. (1991) "Bayesian Methods and Entropy in Economics and Econometrics", in *Maximum Entropy and Bayesian Methods*, eds. W. T. Grandy, Jr. and L. H. Schick, 17-31, Netherlands: Kulwer.
- Zellner, A. (1996a) "Bayesian Method of Moment / Instrumental Variable (BMOM/IV) Analysis of Mean and Regression Models", in *Modeling and Prediction: Honoring Seymour Geisser*, J. Lee, W. Johnson, and A. Zellner (eds.), Springer-Verlag.
- Zellner, A. (1996b) "Models, Prior Information, and Bayesian Analysis" *Journal of Econometrics*, 75, 51-68.
- Zellner, A. (1997) "The Bayesian Method of Moments (BMOM), Theory and Applications", in *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, Vol. 12, T. B. Fomby and R. C. Hill (eds.), Greenwich CT: JAI Press, 85,106.
- Zellner, A. and B. Sacks (1996) "Bayesian Method of Moment (BMOM) Analysis of the Multiple Regression Model with Autocorrelated Errors", H.G.B. Alexander Research Foundation, University of Chicago.
- Zellner, A., J. Tobias, and H. K. Ryu (1997) "Bayesian Method of Moments (BMOM) Analysis of Parametric and Semiparametric Regression Models", Manuscript presented at the Fourth World Meeting of the International Society for Bayesian Analysis, Istanbul, Turkey.