Rev.R.Acad.Cienc.Exact.Fis.Nat. (Esp)
Vol. 93, N.º 3, pp 381-386, 1999
Monográfico: Métodos Bayesianos en las Ciencias

MAXIMALLY INFORMATIVE STATISTICS

(Bayesian inference/Kullback-Leibler distance/sufficient statistics/mutual information/variational methods)

DAVID R. WOLF[1] AND EDWARD I. GEORGE[2]

[1] PO 8308, Austin, TX 78713-8308. drwolf@realtime.net.
[2] Dept of MSIS, University of Texas. Austin TX 78712-1175, USA. ed.george@bus.utexas.com

ABSTRACT

In this paper we propose a Bayesian, information theoretic approach to dimensionality reduction. This approach is formulated as a variational principle on mutual information, and seamlessly addresses the notions of sufficiency, relevance, and representation. Maximally informative statistics are shown to minimize a Kullback-Leibler distance between posterior distributions. To illustrate the approach, we derive the maximally informative one dimensional statistic for a random sample from a Cauchy distribution.

RESUMEN

Estadísticos máximo-informativos

En este trabajo proponemos una aproximación Bayesiana, basada en la teoría de la información, al problema de la reducción de la dimensionalidad. El procedimiento se formaliza como un principio variacional sobre la información mutua, y permite un tratamiento adecuado de las nociones de suficiencia, relevancia y representatividad. Demostramos que los estadísticos máximo-informativos minimizan una distancia de Kullback-Leibler entre distribuciones finales. Para ejemplificar el procedimiento, desarrollamos el estadístico unidimensional máximo-informativo correspondiente a una muestra aleatoria de una distribución de Cauchy.

---

[1] David R. Wolf is a Physicist and Independent Researcher, PO 8308, Austin, TX 78713-8308, wolf@lanl.gov. Edward I. George is the Ed and Molly Smith Chair and Professor of Statistics, Department of MSIS, University of Texas, Austin, TX 78712-1175, egcorge@mail.utexas.edu. All correspondence should be addressed to Dr. Wolf. This work was suported by NASA Center for Excellence in Information Technology contract NAS-214217, NSF grand DMS-98.03756 and Texas ARP grants 003658.452 and 003658.690.

1. INTRODUCTION

Dimensionality reduction is a fundamental goal of tatistical science. In a modeling context, this is often facilitated by estimating a low dimensional quantity of interest. For example, suppose the quantities of interest are the labels of a classification of photographs of objects; of trees, children, etc. The data are the photographs, and the goal is to infer which of the several classes have been presented. In this case the data space often has dimension on the order of $>10^6$, while the parameter space is a small discrete set of labels each having much lower dimension. A low dimensional summary of the photograph is then obtained as the estimate of the classification of the photograph.

In this paper, we propose a Bayesian approach to dimensionality reduction based on maximizing the mutual information between a statistic and a quantity of interest. This approach is formulated as a variational principle on mutual information, and seamlessly addresses the notions of sufficiency, relevance, and representation. We refer to statistics which maximize this mutual information as *maximally informative* (MI) statistics. Such statistics are shown to minimize a Kullback-Leibler distance between posterior distributions.

The mutual information between a statistic and a quantity of interest is defined in Section 2. The mutual information based variational principle for MI statistics is utilized in Section 3 to derive non-variational derivative forms of the principle. In Section 4 several properties of MI statistics are derived. The important result of this section is that MI statistics provide a generalization of the notion of sufficiency, because they are sensible both when they are not sufficient statistics, and when lower-than-data-dimension sufficient statistics do not exit. In Section 5 we present the result that in inference the Kullback-Leibler (KL) distance is properly a functional of posterior distributions. There we find MI statis-

tics at functional minima of a KL distance based on posterior distributions of the parameter of interest. The arguments made here suggest that the KL distance derived here is preferred to a maximum relative entropy distance, a fact which is not discussed in, for example, Kullback [1959] or Shore and Johnson (1979), and numerous others. In Section 6 the MI static for the location parameter of the Gaussian distribution is derived, and shown to be the expected result, since in this case a one-dimensional sufficient statistic exists. In Section 7 we find a one-dimensional MI statistic for the Cauchy distribution, where a sufficiency reduction does not exist. In Section 8 we discuss approximating the posterior distribution as a Gaussian and apply this technique to show that the MI statistics are then Bayes' estimators of the mean and standard deviation. There a contrast of the approximate MI inference approach with the Maximum Entropy method is made, and it is shown that although they agree for Gaussian likelihoods, they disagree for other distributions, with simplicity arguing in favor of the MI statistics.

2. THE MUTUAL INFORMATION BETWEEN A STATISTIC AND A QUANTITY OF INTEREST

Let the data $x \in X$ be drawn according to a parameterized distribution $P(x \mid \theta)$, with $\theta \in \Theta$, the parameter space. θ itself is distributed according to the prior $P(\theta)$. The marginal distribution of x is obtained from $P(x) = \int P(x \mid \theta)P(\theta)\, d\theta$, and the posterior of θ given x is obtained from Bayes Theorem as

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)} \tag{1}$$

The quantity of interest $q = \xi_Q(\theta)$ will be a function of θ, a mapping from the parameter space Θ into some Q, $\xi_Q(\cdot)$:$\Theta \rightarrow Q$. It will be useful to use the Dirac delta-function $\delta(\cdot)$ to represent the distribution of q as

$$\begin{aligned} P(q \mid \theta) &= P(\{\theta\colon q = \xi_Q(\theta)\} \mid \theta) \\ &= \delta\big(q - \xi_Q(\theta)\big) \qquad (2) \\ &= \prod_{\delta=1}^{k_q} \delta\big(q_i - \xi_{Q,i}(\theta)\big), \qquad (3) \end{aligned}$$

where $\delta\big(z(\cdot)\big) = \Pi_i \delta\big(z_i(\cdot)\big)$. Note that (2) may be seen directly by using Bayes' theorem to expand $P(q, \theta)$ as $P(q \mid \theta)P(\theta)$, integrating that over q, which must produce $P(\theta)$, and noting that because the support of $P(q \mid \theta)$ is the unique q such that $q = \xi_Q(\theta)$ (θ is specified), $P(q \mid \theta)$ must therefore be the Dirac delta function. The distribution of q given the data x, may be written using (1) and (3) as

$$P(q \mid x) = \int P(q \mid \theta)P(\theta \mid x)\, d\theta \tag{4}$$

A statistic $r = \xi_R(x)$ will be a function of x, a mapping from the data space X into some R, $\xi_R(\cdot)$: $X \rightarrow R$. Again using the delta notation, the distribution of the statistic is represented as

$$\begin{aligned} P(r \mid x) &= \delta\big(r - \xi_R(x)\big) \qquad (5) \\ &= \prod_{i=1}^{k_r} \delta\big(r_i - \xi_{R,i}(x)\big) \qquad (6) \end{aligned}$$

The joint distribution of the statistc r and the quantity of interest, q, conditioned on the data x is

$$P(r, q \mid x) = P(r \mid x)P(q \mid x) \tag{7}$$

(since $r = \xi_R(x)$ is specified once x is known, making $P(r \mid x, q) = P(r \mid x)$), and the unconditional joint distribution is

$$P(r, q) = \int P(r \mid x)P(q \mid x)P(x)\, dx \tag{8}$$

Finally, we define the mutual information between a statistic and a quantity of intereset as

$$M\big(\xi_R(\cdot), \xi_Q(\cdot)\big) = \iint P(r, q) \log\left(\frac{P(r, q)}{P(r)P(q)}\right) dq\, dr \tag{9}$$

The mutual information is the Kullback-Leibler distance between the joint distribution $P(r, q)$ and the marginal product $P(r)P(q)$ corresponding to independence between r and q.

3. MI STATISTICS AND THE VARIATIONAL PRINCIPLE

We are now ready to define a maximally informative (MI) statistic.

Let $S = \{\xi_R(\cdot)\}$ be a set of statistics under consideration. A MI statistic for a quantity of interest $\xi_Q(\cdot)$ is any statistic $\xi_R(\cdot)$ from S maximizing the mutual information $M\big(\xi_R(\cdot), \xi_Q(\cdot)\big)$ between the statistic and the quantity of interest.

The following variational principle can be used to obtain an MI statistic. Let $\dfrac{\delta}{\delta f(\cdot)}$ denote the functional derivative with respect to $f(\cdot)$.

Choose $\xi_R(\cdot)$ from S such that $\dfrac{\delta M\big(\xi_R(\cdot), \xi_Q(\cdot)\big)}{\delta \xi_R(\cdot)} = 0$ and $\dfrac{\delta^2 M\big(\xi_R(\cdot), \xi_Q(\cdot)\big)}{\delta \xi_R(\cdot)^2}$ is negative semidefinite, i.e. so that $\xi_R(\cdot)$ maximizes the information between itself and $\xi_Q(\cdot)$, the

quantity of interest. If possible, choose the global maximum.

Note that MI statistics in S may occur on the boundary of S. This may be a case of interest, which occurs when constraints are imposed on the statistics, and can be handled with a trivial modification. Note that the space S of statistics can be constrained to contain only low-dimensional statistics to force a dimesionality reduction of the data.

We now demonstrate the variational principle for MI statistics. The argument proceeds by varying (see, for example, Arfken (1985) for the variational calculus) the mutual information of (9) with respect to the statistic function $\xi_R(\cdot)$ of dimension k_r, i.e. $\xi_R(\cdot) = (\xi_{r,1}(\cdot), ..., \xi_{r,k_r}(\cdot))$. We now proceed to substitute $\xi_R(x) = \xi^0_R(x) + \varepsilon\eta(x)$ in (9), and take the derivative with respecto to ε.

Assuming appropriate regularity conditions, we have

$$\partial_\varepsilon M(\xi_R(\cdot), \xi_Q(\cdot)) = \iint \left[\partial_\varepsilon P(r, q) \log\left(\frac{P(r, q)}{P(r)P(q)}\right) + \right.$$

$$\left. + P(r)\partial_\varepsilon P(q \mid r) \right] dq\, dr = \tag{10}$$

$$= \iint \partial_\varepsilon P(r, q) \log\left(\frac{P(r, q)}{P(r)P(q)}\right) dq\, dr, \tag{11}$$

where simplification from (10) to (11) occurs because probability is conserved. Utilizing (7) we find

$$P(r, q) = \int \delta(r - \xi_R(x))P(q \mid x)P(x)\, dx \tag{12}$$

Taking the derivative of (12) with respecto to ε yields

$$\partial_\varepsilon P(r, q) = \sum_{j=1}^{k_r} \int \delta'(r_j - \xi_{R,j}(x))\eta_j(x) \times$$

$$\times \prod_{i \neq j} \delta(r_i - \xi_{R,i}(x))P(q \mid x)P(x)\, dx \tag{13}$$

Note that because η is arbitrary, we may choose it to simplify as needed.

We proceed by considering k_r choices of η. Label the choices by $m \in \{1, ..., k_r\}$, and on choice m take the components of η as follows:

$$\eta_\ell(x) = \delta(x - x_c), \ (\ell = m) \tag{14}$$

$$n_\ell(x) = 0, \quad (\ell \neq m) \tag{15}$$

where x_c is any data point we may choose. The condition that the mutual information is extremal then becomes the statement that for all x_c and $i \in \{1, ..., k_r\}$.

$$\partial_\varepsilon M(\xi_R(\cdot), \xi_Q(\cdot))\,|_{\varepsilon=0} = 0 \tag{16}$$

$$= \iint \delta'(r_i - \xi^0_{R,i}(x_c)) \prod_{i \neq j} \delta(r_i - \xi^0_{R,i}(x_c)) \times$$

$$\times P(q \mid x_c) \log\left(\frac{P(r, q)}{P(r)P(q)}\right) dq\, dr \tag{17}$$

Integrating (17) by parts with respecto to r (dropping both the «0» superscript and subscript «c», since there is no distinction to be made at this point) yields the condition that for all x

$$\int P(q \mid x)\, \partial_r \log\left(\frac{P(r, q)}{P(r)P(q)}\right)\Bigg|_{r=\xi_R(x)} dq = 0 \tag{18}$$

where derivatives with respect to vectors are gradients (vectors of derivatives). The form from which the theorems of the next section are proven, is found by rewriting (18) as

$$\int \frac{P(q \mid x)}{P(q \mid r)}\, \partial_r P(q \mid r)|_{r=\xi_R(x)}\, dq = 0 \tag{19}$$

4. MI STATISTICS AND SUFFICIENCY

Now we prove several important properties concerning MI statistics. The first property is the intuitively obvious property that *data is a MI statistic*. The second property is that *any sufficient statistic is a MI statistic*. Finally, we note that *MI statistics are not necessarily sufficient statistics.*

Theorem 1. *Any* 1-1 *function of data is a MI statistic of the quantity of interest.*

Proof: Let $\xi_R(\cdot)$ be the identity so that $\xi_R(x) = x$ in (19). The fraction in that equation is then 1, and the derivative integrates to zero because probability is conserved. Having $\xi_R(x)$ any invertible function changes nothing since it determines x.

Theorem 2. *Any sufficient statistic for the quantity of interest is a MI statistic of the quantity of interest.*

Proof: Note that using the definition of $\xi_R(x)$ being a sufficient statistic the ratio in (19) is one - the posterior distribution of the quantity of interest given the data x is the same as the posterior distribution of the quantity of interest given the sufficient statistic $\xi_R(x)$. The derivative then integrates to zero because probability is conserved.

(Note that in both Theorems 1 and 2 the Hessian condition of the MI inference variational principles is easily established since then the extremum of the mutual information is easily seen to be a local maximum. Otherwise, one must check the convexity.)

Although it is true that any sufficient statistic is a MI statistic, the converse is false. In problems where a sufficiency reduction does not exist, there will exist lower dimensional MI statistics. Thus, the class of maximally informative statistics contains the sufficient statistics, but is broader. MI statistics need not provide all of the available information about the underlying quantity of interest. For example, as we show in Section 7, such a one-dimensional MI statistic can be obtained for the Cauchy distribution where a sufficiency reduction is unavailable. In a sense, MI statistics seamlessly address *relevance* to the consumer of the information because it is about some *relevant quantity of interest* that MI statistics are maximally informative.

5. MI STATISTICS AND THE KL DISTANCE

Equation (19) may be rewritten as

$$\partial_r\left[\int P(q\,|\,x)\,\log\left(\frac{P(q\,|\,x)}{P(q\,|\,r)}\right)dq\right]\Bigg|_{r=\xi_R(x)} = 0 \quad (20)$$

which, along with the curvature condition, states that

Theorem 3. *The Kullback-Leibler distance between the posterior distribution conditioned on the statistic and the posterior distribution conditioned on the data is minimized by a MI statistic.*

Again, note that MI statistics for the quantity of interest are generally not sufficient statistics for the quantity of interest. Indeed, rather than making the Kullback-Leibler distance zero, as in the case of sufficient statistics, MI statistics are found at local minima of the Kullback-Liebler distance-viewed as a functional of the statistic. This demonstrates how the approach of this paper generalizes that performed by Lindley (1961).

6. MI STATISTICS FOR THE GAUSSIAN DISTRIBUTION

This section details the inference of the one-dimensional MI statistic for the one-dimensional Gaussian distribution. We take the position parameter of the Gaussian to be q, and the goal is to find $\xi_R(x)$ so that (19) holds. From there note that the calculation of $P(q\,|\,r)$ and $P(q\,|\,x)$ is necessary, and by Bayes' theorem therefore it is necessary to find $P(r\,|\,q)$, which may be written as

$$\begin{aligned} P(r\,|\,q) &= \int P(r\,|\,q,\,x)P(x\,|\,q)\,dx \\ &= \int P(r\,|\,x)P(x\,|\,q)\,dx \\ &= \int \delta(r-\xi_R(x))\prod_{i=1}^{N}\frac{e^{-(x_i-q)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \end{aligned} \quad (21)$$

The ansatz $\xi_R(x) = \sum_{i=1}^{N}\lambda_i x_i$ is useful (and not restrictive since the $\lambda_i's$ are implicitly only restricted to be functions of x), and making the changes of variables $y_i = \lambda_i x_i$ followed by $u_i = \lambda_i q$ in (21) yields a form which may immediately be recognized as the convolution of N Gaussians with means $\mu_i = \lambda_i q$ and standard deviations $\sigma_i = \lambda_i\sigma$ respectively,

$$P(r\,|\,q) = \int \delta\left(r - \sum_{i=1}^{N}\lambda_i q - \sum_{i=1}^{N}u_i\right)\prod_{i=1}^{N}\frac{e^{-u_i^2/2(\sigma\lambda_i)^2}}{\sqrt{2\pi}(\lambda_i\sigma)}\,du. \quad (22)$$

This has the solution

$$P(r\,|\,q) = \phi(0,\,\sigma')\left(r - \sum_{i=1}^{N}\lambda_i q\right) \quad (23)$$

where $\sigma' = \sigma\sqrt{\sum_{i=1}^{N}\lambda_i^2}$ and ϕ is the Gaussian density

$$\phi(\mu,\,\sigma)\,(z) = \frac{1}{\sqrt{2\pi}\sigma}\,e^{-(z-u)^2/2\sigma^2}.$$

Finally, inserting this result into Bayes' theorem with uniform prior to find the posterior distribution of q conditioned on r yields

$$P(q\,|\,r) = S\phi(0,\,\sigma')\left(r - \sum_{i=1}^{N}\lambda_i q\right) \quad (24)$$

where $S = \sum_{i=1}^{N}\lambda_i$.

The calculation for $P(q\,|\,x)$ is similar with the result is that

$$P(q\,|\,x) = \phi\left(\bar{x},\,\frac{\sigma}{\sqrt{N}}\right)(q) \quad (25)$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i$. From the forms of (24) and (25) it is clear that not only will the integrand of (20) (that equation equivalent to (19)) be minimized, but that it will be zero, if all $\lambda_i = 1/N$ is chosen. This of course is the expected result since $\xi_R = \Sigma_{i=1}^{N}x_i/N$ is a sufficient statistic for q when σ is known.

Alternatively, to satisfy that the calculation indicated in (19) is successful at finding the expected result, continue by taking (24) and (25) and substituting them into (19) to find after some simplification the equation which must be satisfied by ξ_R

$$0 = \left[\int \left(r - q \sum_{i=1}^{N} \lambda_i \right) e^{-(\bar{x}-q)^2/2(\sigma/(\sqrt{n})^2)} dq \right] \Bigg|_{r=\xi_R(x)} . \quad (26)$$

This has the unique solution $\xi_R(x) = \bar{x}$ when the arbitrary scale of the inferred statistic is fixed by setting $1 = \Sigma_{i=1}^{N} \lambda_i$. To conclude this section, the procedure culminating in (20) or (19) of finding MI statistics has been shown to produce the expected known result for the Gaussian case. The next section approaches the Cauchy distribution case for lower than data dimension statistics, where there is no sufficient statistic available and the result is novel.

7. MI STATISTICS FOR THE CAUCHY DISTRIBUTION

This section outlines the inference of the one-dimensional MI statistic for the one dimensional Cauchy distribution. The detailed steps may be taken similarly to those of the last section but taking the Cauchy distribution instead of the Gaussian distribution. Take the position parameter of the Cauchy to be q, and the the goal is to find $\xi_R(x)$ so that (19) holds. As in the last section it is necessary to determine both $P(q \mid r)$ and $P(q \mid x)$. Assuming the same ansatz that $\xi_R(x) = \Sigma_i \lambda_i x_i$, the necessary convolutions may be carried out with the use of the Fourier convolution theorem, with the results that

$$P(r \mid q) = \frac{S}{\pi\left(S^2 + (r - qS)^2\right)}, \quad (27)$$

$$P(q \mid r) = \frac{S}{\pi\left(S^2 + (r - qS)^2\right)}, \quad (28)$$

and

$$P(q \mid x) \propto \prod_{i=1}^{n} \frac{1}{\pi\left(1 + (x_i - q)^2\right)} \quad (29)$$

where $S = \Sigma_{i=1}^{N} \lambda_i$. Substituting (27), (28), and (29) into (19) yields the equation that must be solved for $\xi_R(x)$

$$0 = \left[\int \left(\prod_{i=1}^{n} \frac{1}{\pi\left(1 + (x_i - q)^2\right)} \right) \frac{r/S - q}{1 + (r/S - q)^2} dq \right] \Bigg|_{r=\xi_R(x)} . \quad (30)$$

Rewriting this equation in more suggestive terms, while taking the scale $S = 1$, gives the result as an implicit equation for $\xi_R(x)$,

$$\xi_R(x) = \int qP\left(q \mid (x, \xi_R(x))\right) dq. \quad (31)$$

The form of the result (31) says that $\xi_R(x)$ is the posterior mean of q given the data *and itself*. This form also suggests that $\xi_R(x)$ could be the posterior mean of q given the data. However, the surprise is that this is not the case, as a check using the posterior moment forms derived in [6] immediately shows. Further, assuming a value for $\xi_R(x)$ on the right-hand side of (31) allows that to be computed in closed form using the results of Wolf (1998). This finally yields that the left-hand side is a rational function of the right-band side, a fixed point equation which may be solved by standard iterative methods. Other checks immediately show that the solution is not the maximum likelihood solution, nor the median.

To conclude this section, the one-dimensional MI statistic for the Cauchy distribution position parameter has been found as the posterior mean of the position parameter of the Cauchy distribution given the data and the MI statistic, and this statistic is different from the Bayes' estimator which is the posterior mean given the data only.

8. APPROXIMATE MI INFERERNCE AND BAYES ESTIMATORS

In many cases of interest, if not in all cases of relevance with high dimensional data, the convolutions that appear similarly to those in (27) etc., will be quite impossible to do in closed form, and probably in a practical sense will even be numerically intractable. However, there is an approach that may be taken which does some harm to a fully rigorous Bayesian approach, but which may be necessary. The idea that is applicable in these cases of difficulty is to directly take $P(q \mid r)$ in (20) to be Gaussian with parameters $r = \xi_R(x) = (\mu(x), \sigma(x))$. The approximate MI approach just outlined is applied below to finding the approximate MI statistics $(\mu(x), \sigma(x))$. The approximate MI approach is then contrasted with an alternative approach using the KL distance inverted from that of (20), one that resembles Maximum Entropy inference. The results of this section hold for any likelihood, as will become apparent.

Take an aribrary one-dimensional parameterized likelihood parameterized by q (i.e. with q the parameter of interest). Parameterize the inferred distribution $P(q \mid r)$ of (20) as

$$P\left(q \mid r = (\mu, \sigma)\right) = \phi(\mu, \sigma)(q) \quad (32)$$

Equations (20) and (32) imply that the MI statistic is

$$\mu = \int qP(q \mid x)\, dq$$

$$\sigma^2 = \int (q - \mu)^2 P(q \mid x)\, dq \quad (33)$$

These quantities are the Bayes' estimators for the mean and standard deviation of the distribution.

If, on the other hand, the inverted from of the KL distance is taken, as it often is in many of the cases we have observed, the statistic μ is

$$\mu = \frac{\int q\phi(\mu, \sigma)(q) \log\left(P(q \mid x)\right) dq}{\int \phi(\mu, \sigma)(q) \log\left(P(q \mid x)\right) dq} \qquad (34)$$

which, along with another highly non-linear equation for σ, is a highly non-linear system to be solved for $r = (\mu, \sigma)$.

Note that when the likelihood $P(x \mid q)$ is Gaussian that these two approximate approaches produce the same statistic, the posterior mean and standard deviation, but for the Cauchy likelihood, for example, this is not the case with the complex nonlinear system needing to be solved. In contrast the approximate MI inferrence technique always produces the posterior Bayes' moment estimators.

The difference between the forms of the approximate MI statistics and the inverted KL statistics appearing in (34) and (35) respectively makes it clear that one needs a good first-principles approach to the KL distance.

9. CONCLUSION

We have formulated the mutual information based variational principle for statistical inference, a fully Bayesian approach to inference, defined MI statistics for a quantity of interest, shown how the principle may be reformulated as a minimal KL distance principle based on posterior distributions, and demonstrated how inference proceeds when sufficient statistics are absent using the Cauchy distribution. Finally, an approximate approach to the inference of MI statistics was discussed, and the relationship of the resulting statistics to Bayes' estimators and the Maximum Entropy version of the same approximation was noted.

10. ACKNOWLEDGEMENTS

Thanks go to the Data Understanding Group at NASA Ames for their lively and interactive critique, friendship, mentoring, and support. This paper was improved by comments from Dr. Jeremy Frank and Hal Duncan, both of NASA.

REFERENCES

1. Arfken, G. (1985). *Mathematical Methods for Physicists.* Academic Press, Inc., London.
2. Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley, NY.
3. Kullback, S. (1959). *Information Theory and Statistics.* John Wiley and Sons, Inc., New York.
4. Lindley, D. V. (1961). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics,* **27,** 986-1005.
5. Shore, J. E. and Johnson, R. V. (1979). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy. *IEEE Transactions on Information Theory,* **26,** 26-37.
6. Wolf, D. R. (1998). Posterior moments of the Cauchy distribution. In *Maximum Entropy and Bayesian Methods,* (W. Voden Linden et. al., eds.) Kluwer Academic, Dordrecht, Netherlands.