The Philippine Statistician
1990, Vol. 39, Nos. 1-2 & 3-4
pages 12-20

MEASURES OF INFORMATION IN PROBABILITY AND STATISTICS

Adolfo M. de Guzman
Associate Professor
Statistical Center
University of the Philippines

Abstract

Applications of information measures in analysis, probability theory and statistics are pointed out. In testing the independence of random vectors under normality, the maximum likelihood criterion is shown to be equivalent to an entropy-based test. Two large sample tests based on entropic loss of information is presented.

Keywords and phrases: information measures, Fisher information, entropy, entropy-based tests.

1. INTRODUCTION

The idea of using measures of information to prove limit theorems is due to Linnik (1959). He gave an information-theoretic proof of the Central Limit Theorem on Lindeberg conditions. Renyi (1960) provides a measure-theoretic proof of a limit theorem for Markov chains. Only recently de Guzman (1989a) showed that Hadamard's Inequality follows almost trivially from the nonnegativity of the Kullback-Leibler information quantity. Jefferson, May and Ravi (1989) suggest the use of entropy to the scaling of some ordinal categorical data. Entropy-based goodness of fit tests have been developed by Vasicek (1976), Dudewicz and Van der Meulen (1981) and Gokhale (1983). In statistical pattern recognition some rules for feature evaluation are derived from information measures (Ben-Bassat, 1982). Srivastava (1973) attempts to extract the "intrinsic" dimensionality of a multivariate data set by exploiting Shannon's information function. In the last fifteen years, Akaike's Information criterion (AIC) which is based on Kullback-Leibler entropy has found successful application in statistical model evaluation problems (Bozdogan, 1987). Hall (1987) shows that in kernel density estimation, if the kernel is chosen appropriately, likelihood cross-validation does result in asymptotic minimization of Kullback-Leibler loss (information measure). Malvestuto (1989) provides a computational method for obtaining the maximum entropy extension of given discrete probability distributions.

In this paper we shall introduce Fisher's information, give a measure-theoretic proof of Hadamard's inequality, and propose an entropy-based statistic for testing independence of two random vectors.

2. FISHER INFORMATION IN EQUIVALENCE/SINGULARITY DICHOTOMIES

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be an experiment i.e. $(\Omega, \mathcal{P})$ is a measurable space and $\mathcal{P}$ a class of probability measure on $(\Omega, \mathcal{F})$. We say that an equivalence/singularity dichotomy holds if $P, Q \in \mathcal{P}$ implies or $P \equiv Q$ or $P \perp Q$.

Kakutani (1946) gave the following interesting example of dichotomy. Let $(\Omega, \mathcal{P}) = (\overset{\infty}{\underset{i=1}{\pi}} \Omega_i, \sigma(\overset{\infty}{\underset{i=1}{\pi}} \mathcal{F}_i))$ where $(\Omega_i, \mathcal{F}_i)$, is a sequence of measurable spaces satisfying Kolmogorov consistency conditions.Let $\{Q_i\}$. be a sequence of measures, where Q_i is on $(\Omega_i, \mathcal{F}_i)$ for all i. If $\mathcal{P} = \{\pi_{i=1}^{\infty} P_i : P_i \equiv Q_i$ for all $i,\}$ Kakutani showed that if $P = \overset{\infty}{\underset{i=1}{\pi}} P_i$ and $\tilde{P} = \overset{\infty}{\underset{i=1}{\pi_i}} P. \in \mathcal{P}$ then $\tilde{P} \equiv P$ or $P \perp \tilde{P}$ with the former holding iff $\sum_{i=1}^{\infty} H^2(\tilde{P}_i, \tilde{P}_i) < \infty$.

Here H is the Hellinger distance and is defined by

$$2\ H^2(P_i, \tilde{P}_i) = \int \surd(f_i - \surd\tilde{f}_i)\ dv_i$$

where $v_i = \tilde{P}_i + P_i$, $\tilde{f}_i \in d\tilde{P}_i/dv_i$ and $f_i \in dP_i/dv_i$.

In the case where $\Omega = R^{\infty}$, $F = B(R^{\infty})$ and P is the set of all Gaussian measures, let $P = \pi_{i=1}^{\infty} N(0,1)$ and $\tilde{P} = \pi^{\infty} N(\mu_i, \sigma_i^2)$. Feldman (1958) and Hajek (1958) showed that $P \equiv \tilde{P}$ if and only if

$$\sum_{i=1}^{\infty} (\mu_i^2 + (1-\sigma_i)^2) < \infty .$$

In case $\sigma_i^2 = 1$, $P \equiv \tilde{P}$ if and only if

$\mu_i \in l^2$, the space of square summable sequences.

What other probability measures on B(R) besides N(0,1) satisfy the property? Shepp (1965) gave the following answer: if P is a probability measure in (R,B(R) and P_t is the translate of P both then $\pi_i^{\infty} P \equiv \pi_{i=1}^{\infty} P_t$ for all $t_i \in l^2$ if and only if $P \equiv \lambda$ (Lebesque measure) and P has finite information measure that is, there exists a locally absolutely continous density f such that

$$\int \frac{(f')^2}{f} \, d\lambda < \infty$$

Thus in a translation invariant experiment finiteness of Fisher information as defined above is a necessary and sufficient condition for an l^2-type dichotomy

3. AN INFORMATION THEORETIC PROOF OF HADAMARD'S INEQUALITY

Theorem (Hadamard's inequality) If A is an nxp real matrix of rank p, then

$$|A^TA| \le \pi_{i=1}^{p} \sum_{j=1}^{n} a_{ij}^2$$

where **A** = (a_{ij}).

Proof. Consider a p-variate random vector $\mathbf{X} = (X_1,\ldots,X_p)'$ which has p-variate normal density with dispersion matrix $\Sigma = \mathbf{A}^T\mathbf{A}$. The entropies $H(\mathbf{X})$ and $H(X_i)$ $i = 1,..,p$ are given by

$$H(\mathbf{X}) = \frac{p}{2} + \frac{p}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| \ , \ H(X_i) = \frac{1}{2} + \frac{1}{2}\log 2\pi$$
$$+ \frac{1}{2}\log \sigma_i^2 \quad \text{where} \quad \sigma_i^2 = \sum_{j=1}^{n} a_{ij}^2 \quad i=1,\ldots,p$$

so that

$$\sum_{i=1}^{p} H(X_i) = \frac{p}{2} + \frac{p}{2}\log 2\pi + \frac{1}{2}\log\left[\prod_{i=1}^{p}\sum_{j=1}^{n} a_{ij}^2\right]$$

The Kullback-Leibler quantity

$$H(\mathbf{X}) - \sum_{i=1}^{p} H(X_i)$$

is nonnegative and the log function is monotonic yield

$$|\Sigma| = |\mathbf{A}^T\mathbf{A}| \le \prod_{i=1}^{p}\sum_{j=1}^{n} a_{ij}^2 \ ,$$

which is Hadamard's inequality. ▮

4. ENTROPY-BASED TESTS

Vasicek (1979) introduced a test on the composite hypothesis of normality based on sample estimate of entropy. The test was shown to be consistent against all alternatives without a singular continuous part. Its asymptotic normality was exhibited by Dudewicz and van der Meulen under the hypothesis of uniformity and also under a special class of alternative densities. A general form of a goodness-of-fit test statistic for families of maximum entropy distributions was given by Gokhale (1983). The proposed test is shown consistent against an appropriate class of alternatives and simulation and Monte Carlo results show favorable comparison with other goodness-of-fit tests.

To test the independence of two random vectors, de Guzman (1989b) showed that a test based on entropic loss of information is equivalent to the test based on the likelihood ratio criterion under normality assumptions.

Let $X\alpha$, $\alpha = 1,\ldots,N$ be a random sample from $n(\mu,\Sigma)$ and suppose $X' = (X^{(1)\prime}\ X^{(2)\prime})'$ and let μ and Σ be partitioned correspondingly as

$$\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Also, let

$S = \Sigma\ (X_\alpha - X)(X\alpha - X)'/N$ and $A = NS$ with corresponding

partitions $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$ and $A = \begin{bmatrix} A_{11} & a_{12} \\ A_{21} & A_{22} \end{bmatrix}$

Consider the null hypothesis H_o: $X^{(1)}$ and $X^{(2)}$ are independent. Loss of information for discarding $X^{(2)}$ is given by

$$L(X) = H(X) - H(X^{(1)}) = c_{p-m} + \frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_{11}|}$$

under normality assumption.

The null hypothesis is equivalent to

H_o: $\Sigma_{12} = \Sigma_{21} = 0$

so that $|\Sigma| = |\Sigma_{11}||\Sigma_{22}|$. Hence $L(X|H_o) = c_{p-m} + \dfrac{1}{2}\log|\Sigma_{22}|$.

Hence

$$W = L(X) - L(X|H_o) = \frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_{11}||\Sigma_{22}|}$$

We therefore have the following theorem.

THEOREM: For testing the null hypothesis Ho: $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent, the entropy based test statistic

$$W = \frac{1}{2} \log \frac{|\mathbf{A}|}{|\mathbf{A}_{11}|\,|\mathbf{A}_{22}|}$$

is equivalent to the likelihood criterion when $\mathbf{X}$ is multivariate normal.

We now remove the assumption that $\mathbf{X}$ is multivariate normal and construct a nonparametric large sample test based on entropic loss of information. Let $\mathbf{X}$ have density f and suppose f_n is an estimator of f based on the observations $\mathbf{X}_1$, $\mathbf{X}_2, \ldots, \mathbf{X}_n$.

Let $\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix}$ with g, h as the densities of $\mathbf{Y}$ and $\mathbf{Z}$

$$\text{Then } L(\mathbf{X}) = -\int f(\mathbf{x}) \log f(\mathbf{x}) - \int g(\mathbf{y}) \log g(\mathbf{y})$$

and under the null hypotheses Ho: $\mathbf{Y}$ and $\mathbf{X}$ are independent

$$L(\mathbf{X}|\mathrm{Ho}) = -\int g(y)(z)\log[g(y)h(z)]dydz - \int |g(y) \log(y)dx$$

$$= -\int g(y)\log g(y)dy - \int h(z) \log h(z)dz$$

$$+ \int g(y)\log g(y)\, dy = -\int h(z) \log h(z)dz,$$

so that

$$L(X)-L(\mathbf{X}|Ho) = -\int f(\mathbf{x})\log f(\mathbf{x})\,d\mathbf{x} + \int g(\mathbf{y})\log g(\mathbf{y})\,d\mathbf{y}$$

$$+ \int h(z)|\log h(z)dz = H(\mathbf{X}) - H(\mathbf{Y}) - H(\mathbf{Z}).$$

Let f_n, g_n and h_n be density estimates of f,g and h respectively and let

$$W = \hat{H}(\mathbf{X}) - \hat{H}(\mathbf{Y}) - \hat{H}(\mathbf{Z}) = -\int f_n(\mathbf{x})\log f_n(\mathbf{x})\,d\mathbf{x}$$

$$+ \int g_n(\mathbf{y})\log g_n(\mathbf{y})d\mathbf{y} + \int h_n(\mathbf{z})\log h_n(\mathbf{z})d\mathbf{z})$$

If the null hypothesis Ho is true, W should be small. If the density estimates f_n, g_n and h_n are suitably chosen, W can be shown to be normal. Consider $H(\mathbf{X})/\{H(\mathbf{Z}) + H(\mathbf{Y})\}$. Under Ho, $H(\mathbf{X})/H(\mathbf{Z}) + H(\mathbf{Y})\}= 1$. Therefore the problem reduces to testing the null hypothesis Ho: $H(\mathbf{X})/\{H(\mathbf{Z}) + H(\mathbf{Y})\} = 1$ against the alternative A: $H(\mathbf{X})/\{H(\mathbf{Z}) + H(\mathbf{Y})\} < 1$. This suggest the test statistic

$$V = \hat{H}(\mathbf{X})/\{\hat{H}(\mathbf{Z}) + \hat{H}(\mathbf{Y})\}.$$

The asymptotic distribution of V still has to be worked out.

References

Ben-Bassat M. (1982). Use of distance measures, information measures and error bounds in feature evaluation. In: P.R. Krisnaiah and L.N. Kanal. *Handbook of Statistics* **2** North Holland Publishing Company 773-791.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extension. *Psychemetrika* **50** 3 345-370.

de Guzman, A. M. (1989a). Dimension, dimension reduction and loss of information. Dissertation submitted to the University of the Philippines.

__________________, (1986b). An information theoretic proof of Hadamards' inequality. *Matimyas Matematika* (to appear).

Dudewicz, E. J. and van der Meulen (1981). Entropy based tests of uniformity. Journal of the ASA **76**, 967-974.

Gokhale, D.V. (1983). On entropy based goodness-of-fit Test. *Computational Statistics and Data Analysis* **1** 157-165.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics.* **15** **4** 1491-1519.

Jefferson, R. R., May, J.H. and Ravi, N. (1989). An entropy approach to the scaling of ordinal categorical data. Psychometrika **54** 2 203-215.

Linnik, Yu V. (1959). An information theoretical proof of the central limit theorem on Lindeberg conditions. *Teor. Veroyatnost. i Primenen,* **4** 311-321. (In Russian).

Malvestuto, F.M. (1989). Computing the maximum-entropy extension of a given discrete probability distribution. *Computational Statistics and Data Analyses.* **83** 299-311.

Neymann, J. (1960). *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* **1**, 547-561.

Renyi, A. (1970). *Foundations of probability.* Holden-Day Inc. San Francisco.

Shore, J.E. et al (1984). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy. *IEEE Information Theory* **26** 26-37.

Srivastava, J.N. (1973). An information function approach to dimensionality analysis and curved maniforld clustering. In: *Multivariate Analysis III*. Academic Press, Inc.

Thelen, B.J. (1989). Fisher information and in equivalence/ contiguity. *The Annals of Probability* **17** 4 1664-1690.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society*. Series B **38** 54-59.