HUITIEME COLLOQUE SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 1er au 5 JUIN 1981



# APPLICATION of MAXIMUM ENTROPY PRINCIPLE to MULTIDIMENSIONAL STATISTICS

Dr. Bülent SANKUR

Boğaziçi University , Department of Electrical Engineering, P.K. 2 Bebek, İstanbul - TURKEY

## RESUME

La méthode d' "Entropie Maximale " est une formalisme mathématique utilisée pour assigner une fonction de densité de probabilité à une ensemble d'information donnée en fonction de ses moments. Cette approche a été appliquée aux statistiques multidimensionnelles en vue d'obtenir des fonctions de distribution d'ordre n à partir des marginaux d'ordre réduit et des contraintes additionnelles. L'utilisation de la méthode d'Entropie Maximale pour les problèmes de détection est indiquée. Une démonstration simple du problème isopérimetrique inverse pour distributions multidimensionnelles est également fournie dans le texte.

## SUMMARY

The Maximum Entropy (ME) procedure is a mathematical formalism to assign a probability density function corresponding to an information set given in the form of moments. This procedure has been applied to multivariate statistics in obtaining n'th order distribution functions from lower order marginals and additional prior constraints. Applications of ME technique to detection problems are also indicated. A simple proof of the inverse isoperimetric problem for multivariate distributions is given.

## I - INTRODUCTION

The principle of maximum entropy and the principle of minimum cross-entropy are mathematical formalism to solve statistical inference and information processing problems in a remarkable variety of fields. Specifically, these problems are characterized by systems or processes whose state probabilities are not known but information on the probability distribution is given in terms of expectations of various functions of the states. To obtain a complete statistical description of the system one needs to assign the probability distribution function of the states which is consistent with the available information but is otherwise minimally prejudiced with respect to the missing information.

These types of problems arises in a variety of situations. In probability theory and applied statistics one often has measurements of the functional moments of the system states and one desires to find the underlying probability distribution function. In time series analysis and modeling one tries to estimate the power spectrum of a process and the model parameters given a limited number of observations. Other applications are in the fields of traffic networks, reliability analysis queuing theory, production line decision making, group behaviour stock market analysis, radio astronomy, geophysical sounding, data communication, etc.

The common aspect is that, in general, the number of observations and data are considerably smaller then the " degrees of freedom " of the system or the process. In this respect the ME solution can be viewed as an antiprojection operation,that is reconstructing the complete system from limited projection data.

## II - MATHEMATICAL PRELIMINAIRES

Consider first a discrete system with n states and M prior constraints in the form of functional moments:

$$\sum_{i=1}^{n} p_i \, a_j(x_i) = A_j \qquad j = 1, \ldots, M \qquad (1)$$

with, $a_1(x) = 1$ which corresponds to the normalization constraint. These constraints, if consistent, define a set of uncountably infinite number of probability distributions, called admissible distributions. Jaynes [1] has shown that among admissible distributions one should choose the probability distribution that maximizes an entropy functional, the Shannon entropy defined as:

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i \qquad (2)$$

subject to the set of constraints in (1). This subjective assignment of the probability distribution can be shown to be also the distribution that would be realized in the greatest number of ways experimentally. Jaynes qualifies this distribution as " the distribution which is maximally noncommittal with regard to missing information and is the one that agrees with what is known but expresses maximum uncertainty with respect to all other matters and thus leaves a maximum possible of freedom for our final decision to be influenced by the subsequent sample data ". The maximum entropy ( ME ) distribution can be found by considering the objective function

$$J_\lambda(p) = H(p) + \sum_{j=1}^{M} \lambda_j \sum_{i=1}^{n} p_i a_j(x_i) \qquad (3)$$

where the constraints have been appended through the Lagrange multipliers $\{ \lambda_j \}$ . The ME distribution is simply found as:

$$p_i = \exp \{ -\lambda_o - \lambda_1 a_1(x_i) - \ldots - \lambda_M a_M(x_i) \} \qquad (4)$$

$$i = 1, \ldots . n$$

Furthermore the Lagrange multipliers in (4) can be simply found by using the constraint equations in (1). This procedure generalizes to the determination of the probability density functions in a straightforward manner using techniques of calculus of variations [2] . Similarly starting from a set of a priori probabilities, an a posteriori distribution in the light of new information can be assigned by minimizing the cross-entropy functional, defined as

$$H(q, p) = \sum_{j=i}^{n} q_j \log (q_j / p_j) \qquad (5)$$

where q is the a posteriori distribution. In essence in the minimum cross-entropy technique one desires to choose a new set of probabilities, q, which is nearest to the given set of probabilities. In the case when the a priori distribution is uniform the minimum cross-entropy technique becomes the maximum entropy technique.

## III- MULTIDIMENSIONAL DISTRIBUTIONS and ME.

Given an n'th order probability density function (p. d. f.) the lower order marginals can be simply found by straightforward integration. On the other hand the problem of obtaining a p. d. f. given its m'th lower order marginals( $m < n$ ) and some additional moment conditions does not have a unique and explicit solution. [4, 5] . Beckmann has solved the special but important case of deriving a second order p. d. f. given the first order marginals and the correlation coefficient using an infinite series of orthogonal polynomials. It will be shown that the ME formalism provides a much more general solution. Infact let the prior contraints be given as

$$\int_{R^{k_j}} f(x_1, \ldots, x_n) \, g_j(x_1, \ldots, x_{k_j}) \, dx_1 \ldots dx_{k_j}$$

$$= G_j(x_{k_j+1}, \ldots, x_n) \qquad (6)$$

$$j = 1, \ldots, M$$

where $k_j$ is a subscript depending upon j, $\{ g_j(.) \}$ are generic functions of $k_j$ variables and $\{ G_j(.) \}$ are moment functions of $n - k_j - 1$ variables. Note that for $k_j = n$ one obtains the functional moments, while for $g_j(.) \equiv 1$ , $k_j < n$, one has the lower order marginals. The performance index becomes in this case:

$$J_\lambda [ f ] = -\int_{R^n} f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n) dx_1 \ldots$$

$$\ldots dx_n + \sum_{j=1}^{M} \int_{R^n} f(x_1, \ldots, x_n) \, \lambda_j(x_{k_j+1}, \ldots, x_n) \cdot$$

$$\cdot g_j(x_1, \ldots, x_{k_j}) \, dx_1 \ldots dx_n \qquad (7)$$

where $\lambda_j(x_{k_j+1}, \ldots, x_n)$ are the Lagrange multiplier functions.

Let us now consider a number of two dimensional distributions to illustrate this formalism.

Case 1: Assume that only the correlation coefficient is available; in other words, the constraint set is given by:

$$\int\int f(x, y) \, dxdy = 1 \quad \text{and} \quad \int\int xy \, f(x, y) \, dxdy = \rho$$

The ME solution then becomes:

$$f(x, y) = \exp \{ -\lambda_o - \lambda_1 xy \} .$$

This equation has no solution in the intervals $(0, \infty)$ and $(-\infty, \infty)$. On the other hand, for a finite interval e.g., $[0, 1]$ one obtains the value of Lagrange multiplier implicity:

$$f = ( e^{-\lambda_1} + Ein(\lambda_1) - 1) / \lambda_1 Ein(\lambda_1)$$

where $Ein(\lambda_1) = - \sum_{n=1}^{\infty} ((-1)^n \lambda_1^n) / (n.n!)$.

Case 2: Let us assume that the marginal distributions e(x) and d(y) are available in addition to the correlation coefficient. It can be shown that the corresponding ME distribution becomes:

$$f(x,y) = \exp \{ -\lambda_o - \lambda_1(x) - \lambda_2(y) - \lambda_3 xy \} \quad (8)$$

where the Lagrange multiplier functions should satisfy

$$\exp(-\lambda_o - \lambda_2(y)) \int \exp(-\lambda_1(x) - \lambda_3 xy)\, dx = d(y) \quad (8a)$$

$$\exp(-\lambda_o - \lambda_1(x)) \int \exp(-\lambda_2(y) - \lambda_3 xy)\, dy = e(x) \quad (8b)$$

Assuming further that the marginals are identical (i.e, $e(.) \equiv d(.)$) it follows then that $\lambda_1(.) = \lambda_2(.) \triangleq \lambda(.)$ and (8) reduces to a nonlinear integral equation of the Hammerstein type [11] :

$$ch(y) \int e^{-rxy} h(x)dx \doteq e(y) \quad (9)$$

where we have used $h(x) \triangleq \exp(-\lambda(x))$, $c \equiv \exp(-\lambda_o)$, $r = \lambda_3$. Under more general moment conditions but still with symmetrical marginals, (9) remains the same in form with the kernel:

$$k(x, y) = \exp \{ - \sum_{j=1}^{M} \lambda_j a_j(x, y) \}$$

Case 3: A common prior constraint is the conditional expectation which in the most general form can be written as

$$E \{ u(x) / v(y) \} = c(y) \quad (10)$$

where E {.1.} denotes the conditional expectation. The constaint set becomes now

$$\int\int f(x, y)\, dxdy = 1 \qquad \int f(x, y)\, dx = d(y)$$

$$\int f(x, y)\, dx = e(x) \qquad \int u(x)\, f(x / y)\, dx = c(y)$$

and the performance index can be written as:

$$J_\lambda[f] = - \int\int f(x, y) \log f(x, y)\, dxdy + \lambda_o \int\int f(x, y)\, dxdy + \int\int \lambda_1(y)\, f(xy)dxdy + \int\int \lambda_2(x)\, f(x, y)\, dxdy + \int \lambda_3(y) \frac{\int u(x)\, f(x, y)\, dx}{\int f(x, y)\, dx} dy \quad (11)$$

Recalling that $\int u(x)\, f(x, y)\, dx = c(y)d(y)$, the ME distribution becomes:

$$f(x, y) = \exp \{ -\lambda_o - \lambda_1(x) - \lambda_2(y) - \lambda_3(y) \cdot \left[ \frac{u(x) - c(y)}{d(y)} \right] \} \quad (12)$$

where the Lagrange multiplier functions, if f(x, y) is symmetric in its marginals, must satisfy a symmetry condition of the type:

$$\int\int u(x)\, h(x, y)\, dxdy = \int\int c(y)\, h(x, y)\, dxdy \quad (13)$$

where $h(x, y) \triangleq \exp \{ - \lambda_1(x) - u(x)\, \lambda_3(y) / d(y) \}$

Case 4: In certain cases the characteristic function of a process has a recursive structure. For example consider the set of vector autoregressive (AR) processes,

$$\underline{x}_n = \underline{\underline{A}}\, \underline{x}_{n-1} + \underline{u}_n$$

where $\{ \underline{u}_n \}$ is a sequence of independent identically distributed random vectors with a joint p. d. f. $g(\underline{u})$ $\underline{\underline{A}}$ is an nxn matrix and $\{ \underline{x}_n \}$ are the sample vectors of the AR sequence. It can be seen that the ch.f. satisfies the equation

$$\phi_{\underline{x}}(v) = \phi_{\underline{x}}(\underline{\underline{A}}\, \underline{v}) \cdot \phi_{\underline{u}}(\underline{v})$$

The prior constraints in this case becomes:

i) $\int_{R^n} f(\underline{x})\, d\underline{x} = 1$

ii) $f(\underline{x}) = \int_{R^n} f(A^{-1}\underline{z})\, g(\underline{z} - \underline{x})\, d\underline{z}$

ii) other moment constraints.

The resulting ME disturbing using constraints i) and ii) becomes

$$f(\underline{x}) = \exp \{ - \lambda_o - \int \lambda(u)\, g(\underline{\underline{R}}\underline{x} - \underline{u})\, \underline{\underline{R}}\, du - \lambda(x) \} \quad (14)$$

IV - APPLICATIONS to DETECTION PROBLEMS

Consider the detection of a known signal in the presence of noise, i.e., the binary hypothesis testing problem

$$H_0 : r(t) = s(t) + n(t) \qquad (15)$$
$$H_1 : r(t) = n(t) \qquad t \in [0,T].$$

where $n(t)$ is a sample function of an additive noise process and $s(t)$ is the signal component. The Bayes detector can be implemented using the likelihood ratio

$$\Lambda[r(t)] = \frac{f(r(t) / H_1)}{f(r(t) / H_0)} \underset{H_0}{\overset{H}{\gtrless}} \eta$$

where $f(r(t) / H_i) = f_i(r)$ $i = 0,1$ are the p.d.f.'s of the received signal under $H_0$ and $H_1$.

Case 5: Suppose that $f_0(r)$ and $f_1(r)$ are not known a priori but information on the moment conditions only is given. One would like to derive in the ME sense the p. d. f.'s $f_0(r)$ and $f_1(r)$ that satisfy the given moment conditions while maximizing ( minimizing ) the detection probability. The corresponding p. d. f. pair will be the most ( least ) favorable distribution in the class of admissible distributions and upper and lower bounds to the receiver performance could thus be set. The performance index can be written as:

$$J_\lambda(f_1, f_0) = -\int f_0(x)\log f_0(x)dx - \int f_1(x)\log f_1(x)dx$$
$$+ \sum_{i=0}^{m} \lambda_i a_i(x) f_1(x)dx + \sum_{i=0}^{n} \eta_i b_i(x) f_0(x)dx$$
$$+ c_1 \int_{D_1} f_1(x)dx + c_0 \int_{D_0} f_0(x)dx \qquad (16)$$

In (16), $D_1$ and $D_0$ are the decision regions for correct detection under hypotheses $H_1$ and $H_0$, respectively. Furthermore $c_1$ and $c_0$ denote the relative weights associated with decisions of each type. The resulting ME distributions become than

$$f(x/H_1) = \exp\{ - \sum_{i=1}^{m} \lambda_i a_i(x) + c_1 u(x - \gamma) \} \qquad (17a)$$
$$f(x/H_0) = \exp\{ - \sum_{i=1}^{m} \eta_i b_i(x) + c_0 u(\gamma - x) \qquad (17b)$$

where $u(.)$ denotes the unit step function. The probability of detection $P_D$ and false alarm $P_F$ can be calculated in terms of $\gamma$ and the receiver operating characteristics (ROC) can be then evaluated. Inverting the signs of the last two terms in (16) one obtains the least favorable distribution for the detection problem. Furthermore for a detector of Neyman - Pearson type one adds simply one more Lagrange multiplier term, i.e., $c_0$ is substituted with $\eta_{n+1}$. It is interesting to note that for ME distributions the the likelihood ratio has always the form

$$\sum_{i=1}^{n} \lambda_i a_i(x) - \sum_{i=1}^{m} \eta_i b_i(x) \underset{}{\overset{}{\gtrless}} \gamma$$

## V - INVERSE ISOPERIMETRIC PROBLEM:

The problem of finding the probability density function that maximizes the entropy functional subject to prior constraints as in (6) is also referred to as the inverse isoperimetric problem [6] . One would like to establish whether the ME distribution is the unique extremal distribution that satisfies the given constraint equations. The following theorem is obtained using a straightforward generalization of the theorem proven in [6] .

Theorem: Given a p. d. f. of the form

$$f(x) = \exp\{ - \sum_{i=1}^{M} \eta (x_{k_j+1},\ldots,x_n)\, g(x_1,\ldots, x_{k_j})\}$$

the needed constraints so that this p. d. f. is the extremal distribution of the isoperimetric problem are

$$E\{ g_j(x_1,\ldots, x_{k_j})\} = G_j(x_1,\ldots, x_{k_j}) \qquad (18)$$
$$j = 1,\ldots M.$$

Proof: The ME solution under the constraints in (18) is the function given by

$$\tilde{f}(\underline{x}) = \exp\{ - \sum_{i=1}^{M} \theta_i(x_{k_i+1},\ldots, x_n)\, g(x_1,\ldots x_{k_i})\} \qquad (19)$$

where $\{\theta(\underline{x})\}$ denote the set of Lagrange multiplier functions. The theorem will be proven if one can show that:

$$\theta_i(x_{k_i+1},\ldots, x_n) = \eta_i(x_{k_i+1},\ldots, x_n) \qquad i = 1,\ldots M \qquad (20)$$

Consider now

$$G_j(x_1,\ldots, x_{k_j}) = \int g_j(x_1,\ldots, x_{k_j}) \exp\{ - \sum_{i=0}^{M} \theta_i(x_{k_i+1},\ldots, x_n)\, g(x_1,\ldots, x_{k_i})\} \qquad (21)$$

where $\{ G_j(x_1,\ldots, x_{k_j}) \}$ are now functions of $\{ \theta_i(x_{k_i+1},\ldots,x_n) \}$ . Consider now the variation of $G_j(.)$ with respect to $\theta_i(.)$:

$$\frac{\partial G_j(x_1,\ldots, x_{k_j})}{\partial \theta_i(x_{k_i+1},\ldots, x_n)} = -\int g_j(x_1,\ldots, x_{k_j})\, g_i(x_1,\ldots x_{k_i})\, \tilde{f}(\underline{x})\, dx_1\ldots dx_{k_j} \qquad (22)$$

For a stationary point to exist all variations must be equal to zero. However for $i = j$ (22) is always a positive quantity and it follows then that there are no local maxima and (20) is always satisfied.

VI - DISCUSSION and CONCLUSION

Several Problems of interest remain to be investigated with ME distributions:

i) If the physical properties of a problem are sufficiently well known and one knows the p. d. f. up to a parameter set one uses parametric techniques. On the other hand if our a priori information is very limited (e.g. one knows only that the p. d. f.'s are symmetric) one reverts to nonparametric p. d. f. estimation techniques. The degree of " nonparametricness " of the estimation rule depends in essence upon the cardinality of the set from which one tries to select an appropriate p. d. f.

One then wonders where does really the ME procedure stand in the spectrum of p. d. f. estimation techniques ranging from parametric to strictly nonparametric methods. Furthermore if one has to estimate the moments from noisy data, one has to determine how robust the ME procedure is and how fast and according to what criteria it converges to the true distribution [7, 8] . Recall that in the case of noisy data for the moment constraints one could use the mean square criterion, e.g.,

$$J_\lambda[f] = -\int f(x)\log f(x)dx + \sum_{i=1}^{r} c_i\left[\lambda_i \int f(x)a_i(x)dx - A_i\right]^2$$

where $\{ c_i \}$ is a set of weighting coefficients.

Lacoss [9] has shown that nonlinear spectral estimation techniques in general, the ME spectral analysis technique in particular, provide a data-adaptive windowing effect. It can be conjectured then that as one derives the ME distribution starting from raw data one would have a Parzen windowing effect on the data. These conjectures are yet to be tested out.

ii) Jaynes' formalism assigns a p. d. f. using a priori moment constraints. On the other hand in statistical work one often has measured data. One then would desire to determine the " essential moment functions " In other words " the smallest number of moment functions to extract the largest amount of information from data ", where the moments are estimated as, e.g.

$$E\{ g_j(x) \} \tilde{=} \frac{1}{n} \sum_{i=1}^{n} g_j(x_i) \qquad j = 1,\ldots, M. \quad (23)$$

This problem can alternately be viewed as designing an optimal set of observers to discriminate two processes or distributions ( e.g., the true distribution from a uniform distribution ). An analogous problem in the context of discriminating two Gaussian processes has been solved by Kadota and Shepp [10] .

iii) Determination of the ME p. d. f. as in (4), (12), (14) etc. can also be viewed as a nonlinear mapping from the set $\{ G_j(.) \}$ to the set $\{ \lambda_j(.) \}$ . Infact e.g. (1) can be expressed as

$$F_k(\lambda_1,\ldots, \lambda_M) = 0 \qquad k = 1,\ldots, M. \quad (24)$$

Such a formulation lends itself conveniently to numerical root finding algorithms such as the Newton - Raphson technique [12] . Efficient numerical techniques should then be developed for the above problems.

In conclusion, it has been shown that the ME procedure can be applied easily to derive multidimensional distribution functions but the success and usefulness of this technique depends upon the design of the optimal observers as in (23) and the efficient implementation of numerical algorithms.

APPENDIX A: Solution of the integral equations.

The derivation of ME p. d. f.'s necessitate the solution of certain nonlinear integral equations. Let us now consider the illustrative example encountered in (14):

$$ch(y) \int_D e^{-rxy} h(x)dx = g(y) \quad (A.1)$$

$$D = [0,\infty)$$

This particular integral equation can be solved through i) Laplace transform technique ii) Eigenfunction expansions iii) Iterative methods iv) Numerical methods: Galerkin etc.

a) One realizes that in (A.1) one has the Laplace transform expression such that

$$ch(y)\, H(\rho y) = g(y) \quad (A.2)$$

where $H(s) = \int_o^\infty e^{-xs} h(x)dx$.

Unfortunately with this method one obtains an indirect solution, in that instead of solving for h(.) for a given g(.), one finds feasible solution g(.) for each pair $\{ h(.), H(.) \}$

b) For D = [0, ∞), Equation (A.1) is a singular Fredholm integral equation with eigenfunction $\{ x^{a-1} \}$ and $\{ x^{-a} \}$ for $0 < a < 1$, i.e., the eigenvalues assuming a continuum of values. The characteristic solution becomes

$$\phi(x,a) \quad \sqrt{\Gamma(1-a)}\, x^{a-1} + \sqrt{\Gamma(a)}\, x^{-a} \qquad x > 0 \qquad (A.3)$$

where $\Gamma(.)$ denotes the Gamma function. If we now expand $h(x)$ in terms of $\{ \phi(x,a) \}$, one obtains

$$h(x) = \int_0^1 v(a)\, \phi(x,a)\, da \qquad (A.4)$$

$v(.)$ being the expansion function . Using (A.4) in (A.1) and defining the partitions $\{ y_1,\ldots, y_m \}$, $\{ s_1,\ldots, s_n \}$ and $\{ t_1,\ldots, t_n \}$ over the variables y, s, t, respectively, (A.1) can be numerically solved using the set of m nonlinear algebraic equations, i.e.,

$$\sum_{i=1}^{n} \sum_{i=1}^{n} v(t_i) v(s_i)\, K(i, j, k) = g(y_k) \qquad k = 1,\ldots,m \qquad (A.5)$$

where

$$K(i, j, k) \triangleq \sqrt{\frac{\sin\pi s_j}{\pi}}\, \phi(rx_k; s_j)\, \phi(x_k; t_i)$$

For more general types of kernels one uses an appropriate set of orthogonal functions to obtain a system of algebraic equations as in (A.5).

REFERENCES

1 - E.T. Jaynes, " Prior Probabilities ", IEEE Trans. Syst. Science and Cyber., 4, 227 - 241 (1968).

2 - M. Tribus, Rational Descriptions, Decisions and Designs. Pergamon Press, 1969.

3 - J.E. Shore, R.W. Johnson, " Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy ", IEEE Trans. Information Theory, IT-26, 26 - 37 (1980).

4 - W. Feller, An Introduction to Probability Theory and Its Applications. II. New York, Wiley, 1969, pp. 99.

5 - P. Beckmann, Orthogonal Polynomials for Engineers and Physicists. The Golem Press, Colorado, 1973.

6 - J.P. Noonan, N.S. Tzannes, T. Costello, " On the Inverse problem of Entropy Maximizations ", IEEE Trans. Information Theory, 22, 120 - 123 (1976).

7 - J.V. Campenhout, T.M. Cover, " Maximum Entropy and Conditional Probability " Tech. Report 32, Dept. of Statistics, Stanford University, (1978).

8 - D. Kazakos, " On Nonparametric Estimation of Probability Density Functions ", D. Kazakos, P. Kazakos (Ed.) Nonparametric Methods in Communications, Marcel Dekker, N.Y., 1977.

9 - R.T. Lacoss, " Data Adaptive Spectral Analysis Methods ", Geophysics, 36, 661 - 675 (1971).

10- T.T. Kadota, L.A. Shepp, " On the Best Finite Set of Linear Observables for Discriminating Two Gaussian Signals ", IEEE Trans. on Information Theory, 13, 278 - 284 (1976).

11- W. Pogorzelski, Integral Equations and their Applications, Pergamon Press, 1966.

12- Johnson, R.W., " Determining Probability Distributions by Maximum Entropy and Minimum Cross-Entropy " Association for Computing Machinery, 24 - 29 (1979).