Int. Jr. of Mathematical Sciences & Applications
Vol. 7, No. 2, (July-December, 2017)

ISSN No: 2230-9888
www.ijmsa.yolasite.com

APPLICATIONS OF MAXIMUM ENTROPY PRINCIPLE TO PROBABILITY DISTRIBUTION

R.K.Tuli
Department of Mathematics
S. S. M. College, Dinanagar, India
r_kumartuli@yahoo.co.in

Abstract

In the existing literature of information theory, there are many measures of probabilistic entropy which provide applications to different disciplines of Mathematical Sciences. One such discipline is probability theory where maximum entropy principle can successfully be employed. In the present manuscript, we have used this maximum entropy principle for approximating a given probability distribution. The technique has also been illustrated with the help of a numerical example.

Keywords: Probability distribution, Entropy, Uncertainty, Maximum entropy principle, Directed divergence, Moments.

I. INTRODUCTION

It has been realized that the maximization of the entropy in a class of distributions subject to certain constraints has captured an important role in statistical theory. Stepniak [13] gave a maximum entropy characterization for various distributions whereas Athreya [1] used the maximum entropy principle to characterize a large number of discrete and continuous distributions under certain constraints. It is a known fact that the uncertainty is maximum when the outcomes are equally likely and this maximization is achieved by uniform distribution, as it contains the largest amount of uncertainty. It was Jaynes [6] who introduced the principle of maximum entropy and stated that from the set of all probability distributions compatible with one or several mean values of one or several random variables, choose the one that maximizes Shannon's [12] entropy, given by

$$H(P) = -\sum_{i=1}^{n} p_i \ln p_i \qquad (1.1)$$

The two disciplines, that is, MaxEnt and Statistics are interrelated and can in some sense supplement each other and both address themselves to obtain information about the probability density function. In MaxEnt, we want to estimate a probability density function when only a few of its moments are known whereas in Statistics, we obtain it from the random sample. These two main optimization principles have been introduced by Jayne [6] and Kullback [9]. Jayne's [6] MaxEnt aims at maximizing uncertainty subject to given constraints and it restricts its use only to Shannon's [12] measure of uncertainty. But, uncertainty is too deep and complex a concept to be measured by a single measure under all conditions. Thus, there is need for generalized measures of entropy just to extend the scope of their applications for the study of different optimization principles.

Kullback's [9] principle of minimum cross-entropy aims to choose that probability distribution, out of all those that satisfy the given constraints, which is closest to a given a priori probability distribution. However, the principle confines itself to one measure of directed divergence, namely, the one given by Kullback and Leibler [10]. This measure is given by

$$D(P:Q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \qquad (1.2)$$

Again, directed divergence is too deep a concept to be represented by one measure only. In fact, in the literature, a large number of measures of directed divergence are available to measure distances. Some of these may not be useful, but there are others that are well-tailored for applications. These are the generalized measures of directed divergence or cross-entropy, which can be applied for the study of different optimization principles.

Some other measures of entropy and divergence which find tremendous applications in a variety of disciplines are due to Cohen and Merlino [2], Kayal et. al. [7], Guo and Suo [4], Dragmoir et. al. [3], Kowalski et. al. [8], Lin and Tomamichel [11] etc.

In section II, we have given the applications of maximum entropy probability distribution for approximating a given probability distribution. The technique has also been illustrated with the help of a numerical example.

II. MAXIMUM ENTROPY PRINCIPLE IN APPROXIMATING A GIVEN PROBABILITY DISTRIBUTION

In many practical problems arising in the various fields of Operations Research, we do not get simple expressions for the probability distributions. In all these cases, it becomes very difficult to apply these expressions for further mathematical treatment in the manipulation of new results. Thus, it becomes desirable to approximate these probability distributions. The approximating probability distributions should have some common properties with the given distribution. The simplest property is of having some common moments.

There may be an infinite number of distributions with the same first moment as the given distribution. Our interest lies with that probability distribution which is most unbiased and from the theory of maximum entropy principle, we accept only that distribution which has maximum entropy. This will provide our first approximation to the given distribution. This result is based upon the postulate that most probability distributions are either maximum entropy distributions or very nearly so. To find a better approximation, we try to find that maximum entropy probability distribution which has two moments in common with the given probability distribution. As the number of moments go on increasing, we get better and better approximations. We illustrate the above mentioned fact by considering the following numerical example:

Numerical Example

Let us consider the theoretical weighted probability distribution P, given by

i :	0	1	2	3	4	5	6
p_i :	0.3	0.25	0.15	0.1	0.08	0.07	0.05
w_i :	1	2	3	4	5	6	7

Our problem is to find maximum entropy probability distribution (MEPD) with

(I) same mean;

(II) same mean and p_o;

(III) same mean, p_o and p_1; and

(IV) same first two moments

Our purpose is also to find that maximum entropy probability distribution which is closest to the given probability distribution P.

To solve the above problem, we make use of maximum entropy principle by using Havrada and Charvat's [5] weighted entropy of order 2 and our problem becomes:

(I) Maximize Havrada and Charvat's [5] weighted entropy of order 2 given by

$$H(P;W) = 1 - \sum_{i=0}^{6} w_i p_i^2 \tag{2.1}$$

subject to the following set of constraints

$$\text{(i)} \sum_{i=0}^{6} p_i = 1 \tag{2.2}$$

$$\text{(ii)} \sum_{i=0}^{6} i p_i = 1.82 \tag{2.3}$$

The corresponding Lagrangian is given by

$$L = 1 - \sum_{i=0}^{6} w_i p_i^2 - \lambda \left\{ \sum_{i=0}^{6} p_i - 1 \right\} - \mu \left\{ \sum_{i=0}^{6} i p_i - 1.82 \right\}$$

Hence $\frac{\partial L}{\partial p_i} = 0$ gives

$$p_i = -\frac{\lambda + i\mu}{2w_i} \tag{2.4}$$

Applying (2.2), we get

$$(2.593)\lambda + (4.407)\mu = -2 \tag{2.5}$$

Applying (2.3), we get

$$(4.407)\lambda + (16.592)\mu = -3.64 \tag{2.6}$$

From (2.5) and (2.6), we have

$\lambda = -0.726$ and $\mu = -0.027$

With these values of λ and μ, equation (2.4) gives the following set of probability distribution:

$p_0 = 0.3630$, $p_1 = 0.1881$, $p_2 = 0.1299$, $p_3 = 0.1007$, $p_4 = 0.0832$, $p_5 = 0.0716$, $p_6 = 0.0633$

Obviously, $\sum_{i=0}^{6} p_i = 0.9998 \cong 1$

Thus, the first MEPD P_1 is given by

$$P_1 = \{0.3630,\ 0.1881,\ 0.1299, 0.1007,\ 0.0832,\ 0.0716,\ 0.0633\}$$

(II) In this case, our problem is to maximize Havrada and Charvat's [5] weighted entropy under the set of constraints (2.2), (2.3) and $p_o = 0.3$.

Now $\sum_{i=0}^{6} p_i = 1$ gives that

$$p_o + \sum_{i=1}^{6} p_i = 1$$

or $(1.592)\lambda + (4.407)\mu = -1.4$ (2.7)

Applying (2.3), we get

$$(4.407)\lambda + (16.592)\mu = -3.64 \tag{2.8}$$

Equations (2.7) and (2.8) together give

$\lambda = -1.028$ and $\mu = 0.053$

With these values of λ and μ, equation (2.4) gives the following set of probability distribution:

$p_0 = 0.3000$, $p_1 = 0.2437$, $p_2 = 0.1536$, $p_3 = 0.1086$, $p_4 = 0.0816$, $p_5 = 0.0635$, $p_6 = 0.0507$

Obviously, $\sum_{i=0}^{6} p_i = 1.002 \cong 1$

Thus, the second MEPD P_2 is given by

$$P_2 = \{0.3000,\ 0.2437,\ 0.1536, 0.1086,\ 0.0816,\ 0.0635,\ 0.0507\}$$

(III) In this case, our problem is to maximize Havrada and Charvat's [5] weighted entropy under the set of constraints (2.2), (2.3), $p_o = 0.3$ and $p_1 = 0.25$.

Now, $p_o + p_1 + \sum_{i=2}^{6} p_i = 1$

This gives $\sum_{i=2}^{6} p_i = 0.45$

Applying (2.4), we get

$$(1.093)\lambda + (3.907)\mu = -0.9 \tag{2.9}$$

Also (2.3) gives

$$(3.907)\lambda + (16.092)\mu = -3.14 \tag{2.10}$$

Equations (2.9) and (2.10) together give

$\lambda = -0.953$ and $\mu = 0.036$

With these values of λ and μ, equation (2.4) gives the following set of probability distribution:

$p_0 = 0.3000$, $p_1 = 0.2500$, $p_2 = 0.1468$, $p_3 = 0.1056$, $p_4 = 0.0809$, $p_5 = 0.0644$, $p_6 = 0.0526$

Obviously, $\sum_{i=0}^{6} p_i = 0.999 \cong 1$

Thus, the third MEPD P_3 is given by

$$P_3 = \{0.3000,\ 0.2500,\ 0.1468, 0.1056,\ 0.0809,\ 0.0644,\ 0.0526\}$$

(IV) In this case, our problem is to maximize Havrada and Charvat's [5] weighted measure of entropy of order 2 subject to the set of constraints (2.2) and (2.3) along with the additional constraint given by

$$\sum_{i=0}^{6} i^2 p_i = 6.58 \tag{2.11}$$

The corresponding Lagrangian is given by

$$L = 1 - \sum_{i=0}^{6} w_i p_i^2 - \lambda\left\{\sum_{i=0}^{6} p_i - 1\right\} - \mu\left\{\sum_{i=0}^{6} i p_i - 1.82\right\} - w\left\{\sum_{i=0}^{6} i^2 p_i - 6.58\right\}$$

Hence $\frac{\partial L}{\partial p_i} = 0$ gives

$$p_i = -\frac{\lambda + i\mu + i^2 t}{2w_i} \tag{2.12}$$

Applying (2.2), equation (2.12) gives

$$(259)\lambda + (441)\mu + (1659)t = -200 \tag{2.13}$$

Applying (2.3), equation (2.12) gives

$$(441)\lambda + (1659)\mu + (7441)t = -364 \tag{2.14}$$

Applying (2.11), equation (2.12) gives

$$(1659)\lambda + (7441)\mu + (36659)t = -1316 \tag{2.15}$$

We, next solve the equations (2.13), (2.14) and (2.15). In matrix notation, we can write the above equations as $AX = B$

where $A = \begin{bmatrix} 259 & 441 & 1659 \\ 441 & 1659 & 7441 \\ 1659 & 7441 & 36659 \end{bmatrix}$, $X = \begin{bmatrix} \lambda \\ \mu \\ w \end{bmatrix}$ and $B = \begin{bmatrix} -200 \\ -364 \\ -1316 \end{bmatrix}$

The augmented matrix is

$$[A:B] = \begin{bmatrix} 259 & 441 & 1659 & :-200 \\ 441 & 1659 & 7441 & :-364 \\ 1659 & 7441 & 36659 & :-1316 \end{bmatrix}$$

$$\square \begin{bmatrix} 259 & 441 & 1659 & :-200 \\ 0 & 908.1 & 4616.2 & :-23.46 \\ 0 & 0 & 2566.6 & :84.33 \end{bmatrix}$$

The above matrix gives

$t = 0.0328$, $\mu = -0.192$ and $\lambda = -0.655$

With these values of w, λ and μ, equation (2.12) gives the following set of probability distribution:

$p_0 = 0.3275$, $p_1 = 0.2035$, $p_2 = 0.1513$, $p_3 = 0.1169$, $p_4 = 0.0898$, $p_5 = 0.0662$, $p_6 = 0.0447$

Obviously, $\sum_{i=0}^{4} p_i = 0.999 \cong 1$

Thus, the fourth MEPD P_4 is given by

$$P_4 = \{0.3275,\ 0.2035,\ 0.1513, 0.1169,\ \ 0.0898,\ 0.0662,\ \ 0.0447\}$$

Our next aim is to find that maximum entropy probability distribution which is closest to the given probability distribution P. For this purpose, we use Havrada and Charvat's [5] weighted measure of directed divergence of order 2 to determine the approximity of the distributions to P. We know that Havrada and Charvat's [5] weighted directed divergence is given by

$$D^{\alpha}(P:Q,W) = \frac{1}{\alpha - 1}\left[\sum_{i=0}^{6} w_i\ p_i^{\alpha}\ q_i^{1-\alpha}\ - 1\right], \alpha \neq 1, \alpha > 0,$$

Thus for $\alpha = 2$, we have

$$D(P:Q,W) = \sum_{i=0}^{6} w_i \frac{p_i^{\,2}}{q_i} - 1 \tag{2.16}$$

Using (2.16), we get the following results:

$D(P_1 : P,W) = 1.898$, $D(P_2 : P,W) = 1.8539$, $D(P_3 : P,W) = 1.8247$, $D(P_4 : P,W) = 1.5392$

Hence, we observe that the MEPD P_4 is closest to the given probability distribution P.

Note: We also make out the following observations:

(a) Since, P_2 is based upon information about mean and p_0 whereas P_1 is based upon information about mean only, we must expect that

$D(P_2 : P,W) < D(P_1 : P,W)$. In our case, it is found to be true.

(b) Since, P_3 is based upon information about mean, p_0 and p_1, we must expect that

$D(P_3 : P,W) < D(P_2 : P,W) < D(P_1 : P,W)$. In our case, it is found to be true.

(c) Since, P_4 is based upon information about the first two moments, whereas P_1 is based upon information about mean only, we must expect that

$D(P_4 : P,W) < D(P_1 : P,W)$. In our case, it is found to be true.

(d) Since, P_2 is based upon information about mean and p_0 ,whereas P_4 is based upon mean and second moment and

$D(P_4 : P,W) < D(P_2 : P,W)$

Thus, we conclude that P_0 gives less information than the second moment.

REFERENCES

[1] Athreya, K. B., "*Entropy maximization*", ISU Stat. Dept. Tech. Report, IMA preprint, 1994, pp. 1231.

[2] Cohen, E. G. D. and Merlino, R. L., "*Clausius entropy revisited*", Modern Physics Letters B, Vol. 28, 2014, pp.1-5.

[3] Dragomir, S. S., Dragomir, N. M. and Sherwell, D., "*Sharp bounds for the Jensen divergence with applications*", Miskolc Math. Notes, Vol. 5, 2014, pp. 63-85.

[4] Guo, J. L. and Suo, Q., "*Upper entropy axioms and lower entropy axioms*", Annals of Physics, Vol.355, pp. 217-223.

[5] Havrada, J. H. and Charvat, F., "*Quantification methods of classification process:Concept of structural α-entropy*", Kybernetika, Vol. 3, 1967, pp. 30-35.

[6] Jaynes, E. T., "*Information theory and statistical mechanics*", Physical Review, Vol.106, 1957, pp. 620-630.

[7] Kayal, S., Kumar, S. and Vellaisamy, P., "*Estimating the Renyi entropy of several exponential populations*", Braz. J. Probab. Stat., Vol. 29, 2015, pp. 94-111.

[8] Kowalski, A.M., Martin, M.T. and Plastino, A., "*Generalized relative entropies in the classical limit*", Physica A, Vol. 422, 2015, pp.167-174.

[9] Kullback, S., Information Theory and Statistics, Wiley, New York, 1959.

[10] Kullback, S. and Leibler, R. A., "*On information and sufficiency*", Annals of Mathematical Statistics, Vol. 22, 1951, pp. 79-86.

[11] Lin, S. M. and Tomamichel, M., "*Investigating properties of a family of quantum Renyi divergences*", Quantum Inf. Process. Vol. 14, 2015, pp. 1501-1512.

[12] Shannon, C. E., "*A mathematical theory of communication*", Bell System Technical Journal, Vol. 27, 1948, pp. 379-423, 623-659.

[13] Stepniak, C., "*On characterization of the normal law in the Gauss-Markov model*", Sankhya, Vol. 58, 1991, Series-A, pp. 115-117.