Indian J. pure appl. Math., 32(6) : 801-815, June 2001

ON MEASUREMENT OF STOCHASTIC DEPENDENCE IN MULTIVARIATE DATA

D. S. HOODA AND B. K. HOODA

Department of Mathematics & Statistics, CCS HAU, Hisar 125 004, *India*

(*Received* 4 *January* 2000; *after revision* 7 *August* 2000; *accepted* 8 *January* 2001)

In the present communication, concept of generalized measure of dependence has been reviewed and discussed. Based on the concept of mutual information, an alternative generalized measure of dependence has been proposed. The properties of information theoretic index[9] for measuring the dependence among sets of variates from a multivariate normal population have been studied. An attempt has been made to apply the normalized measure of stochastic dependence to the problem of feature extraction in recognizing the patterns of agricultural development in Haryana State.

Key Words : Mutual Information; Stochastic Dependence; Cross Entropy; Information Loss; Feature Extraction

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_p$ be p components of a $p \times 1$ random vector $\boldsymbol{X}$. Let $f(x_1, x_2, \ldots, x_p)$ be the joint probability density function of p components and $f_i(x_i)$, $i = 1, 2, \ldots, p$ be their marginal probability density functions. If $X_1, X_2, \ldots, X_p$ are stochastically independent, then the joint distribution does not give any more information than that is given by p component distributions. But in case the components are dependent, we get additional information about the degree of dependence between p components defined by Kullback[10], Theil and Fiebig[13], Soofi[12] as mutual information which is given by

$$D = \int \ldots \int f(x_1, x_2, \ldots, x_p) \ln \frac{f(x_1, x_2, \ldots, x_p)}{f_1(x_1) f_2(x_2) \ldots f_p(x_p)} dx_1 dx_2 \ldots dx_p \quad \ldots (1.1)$$

This was also called as the strength of structure by Watanabe[14]. The measure (1.1) was used measuring the dependency between random variables by Bozdogan[1] and Harris[3] as alternative to correlation coefficient which has the following weaknesses refer Kapur and Kesavan[8].

(i) If we have stochastic variables, we want single measure of dependence among them. The $p\ (p - 1)/2$ correlation coefficients do not give such a measure and thus are not useful when we need a single well defined measure of dependence.

(ii) If the variates are independent, the correlation coefficient is zero, but if correlation coefficient is zero, it is not necessary that the variates are independent. In case of multivariate Pareto distribution, the correlation coefficient is zero for every pair of variates, yet all the variates are dependent among themselves.

(iii) The dependence between two attributes in a contingency table cannot be expressed in terms of correlation coefficient.

To overcome these weaknesses Kapur and Kesavan[8] suggested an information theoretic measure of stochastic dependence and recently, Kim[9] introduced and studied a new normalized measure of stochastic dependence among sets of random variables and called it a dependence index (DI).

In the present paper we shall unify the measures of stochastic dependence in a single well defined one and shall study its applications in patterns recognition. In section 2 we define a measure of stochastic dependence among multivariate random variables and its normalized forms are considered in section 3. In section 4 we obtain an expression of dependence measure of multivariate normal distribution. The properties of dependence index are studied in section 5. In section 6 we have studied application of the stochastic dependence measure in pattern recognition.

2. MEASURES OF STOCHASTIC DEPENDENCE

If $h_1(x)$ and $h_2(x)$ are two p.d.f.'s, then Kullback-Leibler's measure of cross entropy of $h_1(x)$ from $h_2(x)$ is defined as

$$\boldsymbol{D}(h_1 : h_2) = \int h_1(x) \ln \frac{h_1(x)}{h_2(x)} dx, \qquad \ldots (2.1)$$

where $D(h_i : h_2) \geq 0$ and vanishes if $h_1 = h_2$ for all x. $D(h_1 : h_2)$ is also a convex function of x.

Let $X_1, X_2, \ldots, X_p$ are p random variables with marginal p.d.f.'s $f_i(x_i)$, $i = 1, 2, \ldots, p$ and $f(x_1, x_2, \ldots, x_p)$ be joint p.d.f. of p random variables. If we define $g(x_1, x_2, \ldots, x_p) = \Pi f_i(x_i)$ as the product of marginal p.d.f's of $X_1, X_2, \ldots, X_p$ then Watanabe[14] proposed the following dependence measure

$$\boldsymbol{D}(f:g) = \int \ldots \int f(x_1, x_2, \ldots, x_p) \ln \frac{f(x_1, x_2, \ldots, x_p)}{f_1(x_1) f_2(x_2) \ldots f_p(x_p)} dx_1\, dx_2 \ldots dx_p$$

Assuming that ranges of $X_1, X_2, \ldots, X_p$ are independent of one another, (2.2) can be written

$$= -S + \sum_{i=1}^{p} S_i, \qquad \ldots (2.2)$$

where $S = -\int \ldots \int f(x_1, x_2, \ldots, x_p) \ln f(x_1, x_2, \ldots, x_p)\, dx_1\, dx_2 \ldots dx_p$ is the entropy of the joint distribution of $X_1, X_2, \ldots, X_p$

and $$S_i = -\int f_i(x_i) \log f_i(x_i)\, dx_i$$

is the entropy of the marginal distribution of X_i ($i = 1, 2, \ldots, p$).

If $X_1, X_2, \ldots, X_p$ are statistically independent, then

$$f(x_1, x_2, \ldots, x_p) = f_1(x_1), f_2(x_2) \ldots f_p(x_p).$$

It implies $\boldsymbol{D}(f: g) = 0$, otherwise $\boldsymbol{D}(f: g) > 0$.

Some other important measures of stochastic dependence proposed by Kapur (1985 *a* and *b*) are

$$D_1 = \int f_1(x_1) f_2(x_2) \ldots f_p(x_p) \ln \frac{f_1(x_1) f_2(x_2) \ldots f_p(x_p)}{f(x_1, x_2, \ldots, x_p)} dx_1\, dx_2 \ldots dx_p, \quad \ldots (2.3)$$

$$D_2 = \frac{1}{\alpha - 1} \left\{ \int f^{\alpha}(x_1, x_2, \ldots, x_p) [f_1(x_1) f_2(x_2) \ldots f_p(x_p)]^{1-\alpha} dx_1\, dx_2 \ldots dx_p - 1 \right\},$$

$$\alpha > 0; \alpha \neq 1 \quad \ldots (2.4)$$

and

$$D_3 = \frac{1}{\alpha - 1} \ln \int f^{\alpha}(x_1, x_2, \ldots x_p) [f_1(x_1) f_2(x_2) \ldots f_p(x_p)]^{1-\alpha} dx_1\, dx_2 \ldots dx_p,$$

$$\alpha > 0; \alpha \neq 1 \quad \ldots (2.5)$$

Each of the measures is non-negative and vanishes if and only if variates are independent.

For other measures of independence for discrete distributions and an infinite class of normalized measures for continuous distributions we refer to Kapur and Dhande[7].

3. Normalized Measures of Stochastic Dependence

The various information-theoretic measures discussed above can take values over the interval $[0, \infty)$ and hence there is an obvious need for normalized measures taking values over the interval [0, 1]. Some of the important normalized measures of dependence available in literature are :

$$\bar{r}_0 = \frac{S_1 + S_2 - S}{S_1 + S_2}; \quad \text{Camargo and Israel}^2;$$

$$\bar{D} = \frac{S_1 + S_2 - S}{S}; \quad \text{Rajski}^{11},$$

$$\bar{D}_1 = \frac{S_1 + S_2 + \ldots + S_p - S}{(p-1) S}; \quad \text{Kapur}^6$$

and

$$\bar{D}_2 = \frac{m}{m-1} \frac{S_1 + S_2 + \ldots + S_p - S}{S_1 + S_2 + \ldots + S_p}; \quad \text{Kapur}^6.$$

All these measures lie between 0 and 1 and attain the lower limit zero for independence of the variables and 1 for the perfect functional dependence.

From a property of Kullback-Leibler's measure of directed divergence, the greater the value of $\boldsymbol{D}(f: g)$, the greater would be dependence among random variables. For practical purpose this dependence measure $\boldsymbol{D}(f: g)$ is normalized to the following measures :

(*a*) It can be easily verified that

$$S_i \leq S \text{ for each } i = 1, 2, ..., p. \qquad \text{... (3.1)}$$

It implies

$$\sum_{i=1}^{p} S_i \leq p\,S, \text{ where } p \geq 2 \qquad \text{... (3.2)}$$

and $$\boldsymbol{D}(f:g) = \Sigma\, S_i - S \leq (p-1)\,S$$

so that if $\boldsymbol{D}_1 = \dfrac{\sum_{i=1}^{p} S_i - S}{(p-1)\,S}$ then $0 \leq D_1 \leq 1$

b) Considering $\boldsymbol{R}(f:g) = 1 - \dfrac{\text{mutual information}}{\text{joint information}} = \dfrac{2S - \Sigma S_i}{S} = 2 - \dfrac{\Sigma S_i}{S}$,

we define another normalized measure of dependence :

$$D_2 = \sqrt{(1 - R^2(f:g)} = \sqrt{1 - \left(2 - \frac{\Sigma S_i}{S}\right)^2}. \qquad \text{... (3.3)}$$

Since, $\dfrac{\Sigma S_i}{S} \geq 1$ therefore, $0 \leq D_2 \leq 1$ and $D_2 = 0$ if $\boldsymbol{R}(f:g) = 1$ or if $X_1, X_2, ..., X_p$ are stochastically independent.

$D_2 = 1$ if $\boldsymbol{R}(f; g) = 0$ or if mutual information = joint information i.e. $X_1, X_2, ..., X_p$ are perfectly dependent.

(*c*) On the pattern of Kapur and Dhande[7], Kim[9] defined dependence index as a normalized measure of the generalized mutual information $\boldsymbol{D}(f:g)$ as follows :

$$D_3 = 1 - e^{-D(f:g)} \qquad \text{... (3.4)}$$

For convenience sake $D_3(f:g)$ is denoted as $DI(f:g)$.

Remark 1 : Since $\boldsymbol{D}(f:g) \geq 0$ therefore $0 \leq DI(f:g) \leq 1$.

Remark 2 : $DI(f:g) = 0$ if and only if $f(x)$ is not distinguishable from $g(x)$ or if $\Sigma S_i = S$ i.e., the variales $X_1, X_2 \ldots X_p$ are stochastically independent.

Remark 3 : $\boldsymbol{DI}(f:g)$ approaches to 1 as $\boldsymbol{D}(f:g)$ increases infinitely where the reference distribution $g(x)$ is constrained to be independent among the subsets of variables $X_1, X_2, ..., X_p$.

One can use any of these three normalized measures of dependence among X_i's. However, for practical purpose the dependence index $\boldsymbol{DI}(f:g)$ is advantageous.

4. DEPENDENCE MEASURE FOR MULTIVARIATE NORMAL DISTRIBUTION

The density function for the multivariate normal distribution in subsets of m variates is given by

$$f(X_1, X_2, \ldots, X_m) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left[-(1/2)(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

$$X^T = (X_1, X_2, \ldots, X_m)^T \text{ and } \mu^T = (\mu_1, \mu_2, \ldots, \mu_m)^T$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \cdots & \Sigma_{mm} \end{bmatrix}$$

where $\sum_{ij} (>0) = \text{Cov}(X_i, X_j)$ for $i, j = 1, 2, \ldots, m$, then the random vector X has $Np_i\left(\mu_i, \sum_{ii}\right)$ distribution with marginal joint probability density function

$$g_i\left(x_i/\mu_i, \sum_{ii}\right),\ i = 1, 2, \ldots, m.$$

If we consider $f(x/\mu\,\Sigma)$ as true distribution and $g(x/\mu, \Sigma) = \prod g_i \left(x_i/\mu_i, \sum_{ii}\right)$ as reference distribution, then cross entropy or mutual information is given by

$$D(f:g) = \int f(x/\mu, \Sigma) \ln f(x/\mu, \Sigma)\, dx - \sum_{i=1}^{p} \int g_i(x_i/\mu_i, \mu_i) \ln g_i(x_i/\Sigma_{ii}, \Sigma_{ii})\, dx_i \quad \ldots(4.2)$$

$$= \frac{p}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{p}{2} + \sum_{i=1}^{m} \left[\frac{p_i}{2} \ln 2\pi + \frac{1}{2} \ln \left|\sum_{ii}\right| + \frac{p_i}{2}\right]$$

$$= \frac{1}{2} \sum_{i=1}^{m} \ln \left|\sum_{ii}\right| - \frac{1}{2} \ln |\Sigma| \quad \ldots (4.3)$$

$$= \frac{1}{2} \ln \frac{\prod \left|\sum_{ii}\right|}{|\Sigma|} \quad \ldots (4.4)$$

Hence (4.3) can be expressed

$$D(f:g) = -\frac{1}{2} \ln \frac{|R|}{\prod_{i=1}^{m} |R_{ii}|}, \text{ where } R \text{ is correlation matrix.} \quad \ldots (4.5)$$

Thus eq. (3.4) together with (4.5) gives an expression for the dependence index in terms of correlation coefficient —

$$\boldsymbol{DI}\, f: g) = 1 - \left(\frac{|R|}{\prod_{i=1}^{m} |R_{ii}|} \right)^{1/2}, R_{ii} > 0. \qquad \ldots (4.6)$$

It may be noted that the dependence index becomes zero when x_i, are independent and takes value 1 if $|\Sigma| = 0$ for $\left| \sum_{ii} \right| \neq 0$ or if determinant of correlation matrix of the random vector is zero i.e., $|\boldsymbol{R}| = 0$ for $|\boldsymbol{R}_{ii}| \neq 0$.

5. PROPERTIES OF $\boldsymbol{DI}\,(f: g)$

Property 1 — $DI\,(f: g)$ gives single well-defined measure of dependence among m sets of multivariate normal variates.

$$\boldsymbol{DI}\,(f: g) = 1 - \left(\frac{|\boldsymbol{R}|}{\prod_{i=1}^{m} |\boldsymbol{R}_{ii}|} \right)^{1/2}, \boldsymbol{R}_{ii} > 0 \qquad \ldots (5.1)$$

Eq. (5.1) gives a relation between correlation measures of dependence between normally distributed variates and ***DI***, which is illustrated by the following examples :-

a) For bivariate normal distribution $|\boldsymbol{R}| = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = 1 - \rho^2$

$$\boldsymbol{DI} = 1 - (1 - \rho^2)^{1/2}.$$

b) If $p_1 = 1, p_2 = p - 1$,then

$$\boldsymbol{DI} = 1 - (1 - \rho^2_{12 \ldots\ldots p})^{1/2}, \text{ where } \rho_{12 \ldots\ldots p}$$

denotes multiple correlation between normal variate $\boldsymbol{X_1}$ and other $p - 1$ normal variates.

c) If $p_1 + p_2 = p, p_1 \leq p_2$ and $p_1, p_2 \neq 1$, then

$$\boldsymbol{DI} = 1 - \left(\prod_{j=1}^{p} (1 - \rho^2_{j(c)}) \right)^{1/2},$$

where $\rho_{j(c)}$ is jth largest canonical correlation between two sets of variates $X_1 : p_1 \times 1$ and

$$X_2 : p_2 \times 1.$$

d) If $p_1 = p_2 = \ldots\ldots = p_m = 1$ so that $m = p$

$$DI = 1 - |R|^{1/2}.$$

Property 2 — ***DI*** is invariant with respect to linear transformation within each set of variables.

PROOF : Let $Y_i = C_i X_i + b_i$, where C_i is non-singular matrix of order p_i, X_i is $p_i \times 1$ random vector s.t $\sum p_i = p$, $b_i = p_i \times 1$ with constant components, be a linear transformation for each i = 1, 2, ..., m, then $Y = CX + b$ where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}, \quad C = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & c_m \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} \text{ and } \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Taking expectations on both sides, we have

$$E(Y) = CE(X) = C\mu,$$

$$\Sigma^* = E(Y - EY)(Y - EY)^T = C \Sigma C^T$$

and

$$\Sigma^*_{ii} = E(Y_i - EY_i)(Y_j - EY_j)^T = C_j \sum_{ij} C_j^T$$

Thus, respective distributions of Y and Y_i are $N(\mu^*, \Sigma^*)$ and $N(\mu^*_i, \Sigma^*_{ii})$, i = 1, 2 .... m where $\mu^* = C\mu + b$ and $\mu_i = C_i \mu_i + b_i$.

Under the above transformation the cross entropy in (4.4) is given by

$$D(f^* : g^*) = \frac{1}{2} \ln \frac{\prod_{i=1}^{m} |\Sigma^*_{ii}|}{|\Sigma^*|} = \frac{1}{2} \ln \frac{\prod_{i=1}^{m} |C_i| |\Sigma_{ii}|}{|C|^2 |\Sigma|}$$

$$= \frac{1}{2} \ln \frac{|\Sigma_{ii}|}{|\Sigma|}, \text{ since } |C|^2 = \prod_{i=1}^{m} |C_i| = D(f : g) \qquad \ldots (5.2)$$

In view of (5.2) we can conclude that dependence index of Y is identical with that of X. Hence, ***DI*** is invariant with respect to linear transformation.

6. APPLICATION OF DEPENDENCE MEASURE IN PATTERN RECOGNITION

Pattern recognition is concerned with devising methods to train computers to recognise patterns under noisy and complex situations. If we are given an object, our aim is to assign it the proper class. For example, we are given a sample of wheat, the problem is to assign it to one of a dozen classes

of wheat. Another problem is of selection of a candidate in an interview, the assignment has be made to one of the following classes : Outstanding, good, satisfactory or poor. These are only some simple examples, but there are numerous problems of varying complexity.

Let an object be characterized by a large number of characteristics $X_1, X_2, \ldots, X_n$ which can be represented by a vector

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T \qquad \ldots (6.1)$$

We call random variate $\boldsymbol{X}$ as pattern vector and assume that its probability distribution is known. Suppose, there are $C_1, C_2, \ldots\ldots C_k$ classes and we are to assign given $\boldsymbol{X}$ to one of these classes such that the probability of error of misclassification (PEM) may be the least. In practice, it is inconvenient to handle vectors when n is large, say, 20 or more. As such, before proceeding with the problem of classification we first transform the random vector $\boldsymbol{X}$ from an n-dimensional space to a vector $\boldsymbol{Y}$ in an m-dimensional space, where $m << n$.

We can use a linear transformation

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}, \qquad \ldots (6.2)$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots Y_m)^T$ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ and $\boldsymbol{A}$ is a matrix of order $m \times n$.

Since $m << n$, therefore, the vectors $\boldsymbol{Y}$ are much simpler to handle than the vectors $\boldsymbol{X}$. However, this simplification has been achieved at some cost as we have lost some information because of singular transformation (6.2) and consequently, we have lost some of power of discrimination between vectors which can lead to an increase in PEM.

This problem of choosing $\boldsymbol{A}$ so as to minimize PEM is called the problem of feature extraction or the problem of dimensionality reduction. The vectors $\boldsymbol{Y}$ are called feature vectors. For feature extraction, it is required to express the PEM as a function of $\boldsymbol{A}$, but that is a very arduous job. So instead of minimizing PEM, we minimize some other functions that are closely related to it and are tractable in terms of $\boldsymbol{A}$. In particular, we shall minimize stochastic dependence measure among $Y_1, Y_2, \ldots Y_m$, since the more independent these are, the smaller will be the increase in PEM or we shall choose $\boldsymbol{A}$ such as loss of information is minimum.

Let $f(y_1, y_2, \ldots, y_m)$ be joint density function of vectors $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots. \boldsymbol{Y_m}$ while marginal density function of each vector $\boldsymbol{Y_i}$ is $g_i(y_i)$, for $i = 1, 2, \ldots., m$.

Then

$$\boldsymbol{D}(f:g) = \int \ldots \int f(y_1, y_2, \ldots, y_m) \ln \frac{f(y_1, y_2, \ldots y_m)}{g_1(y_1)\, g_2(y_2) \ldots g_m(y_m)} dy_1\, dy_2 \ldots dy_m$$

$$= \int \ldots \int f(y_1, y_2, \ldots y_m) \ln f(y_1, y_2 \ldots y_m)\, dy_1\, dy_2 \ldots\ldots dy_m$$

$$- \int g_1(y_1) \ln g_1(y_1)\, dy_1 \ldots - \int g_m(y_m) \ln g_m(y_m)\, dy_m$$

$$= \sum_{i=1}^{m} S_i - S. \qquad \ldots (6.3)$$

Next we shall choose A which can minimize D. It can be shown[8] that such a matrix is given by

$$(W_1, W_2, \ldots W_m)^T, \qquad \ldots (6.4)$$

where $W_1, W_2, \ldots W_m)^T$ are Orthonormal Eigen Vectors corresponding to m largest eigen values of the covariance or correlation matrix of pattern vector X.

Illustration — We consider data on following agricultural development indicators for 17 districts of Haryana state (Table I) for the year 1997 :

x_1 : Percentage of gross irrigated area to the total cropped area

x_2 : wheat yield (qt/ha)

x_3 : % of agricultural workers

x_4 : Gross value of agricultural output per capita (rural) at current prices

x_5 : Gross area (%) under commercial crops to the total cropped area

x_6 : Gross value from agriculture per hectare at current prices (0000 Rs.)

TABLE I

District	x_1	x_2	x_3	x_4	x_5	x_6
Ambala	81.1	32.64	40.64	5.573	10.89	27.179
Panchkula	39.2	21.39	29.41	2.589	7.22	16.356
Yamunanagar	83.1	33.03	48.97	6.986	26.23	34.197
Kurukshetra	99.6	37.62	58.79	10.355	9.94	38.156
Kaithal	98.9	36.16	66.95	8.385	4.76	33.259
Karnal	99.1	36.19	56.33	9.439	3.28	34.233
Panipat	99.4	36.81	51.71	4.528	4.4	32.738
Sonipat	93.6	38	42.27	7.216	8.08	27.034
Rohtak	69.6	35.17	54.57	3.234	23.88	17.121
Faridabad	75.8	33.85	35.91	4.15	7.96	22.193
Gurgaon	51.4	36.58	46.64	3.379	21.25	18.29
Rewari	61.2	39.24	47.3	4.059	39.92	18.788
Mahendergarh	46.7	40.69	48.53	4.321	41.5	18.436
Bhiwani	42.5	36.08	58.1	5.798	25.3	17.061
Jind	92.4	37.21	63.16	6.339	21.51	23.477
Hisar	92.4	39.75	64.9	9.747	35.93	27.146
Sirsa	82.0	39.48	62.93	14.119	44.14	29.245

Let $x^T = (x_1, x_2, x_3, x_4, x_5, x_6)$ be the pattern vector of the indicators of agricultural development.

Estimated variance covariance matrix (S) and correlation matrix (R) are

$$S = \begin{bmatrix} 429.868 & 29.853 & 101.683 & 37.022 & -105.909 & 12.452 \\ 29.853 & 18.025 & 27.596 & 5.411 & 27.925 & 0.622 \\ 101.683 & 27.596 & 106.167 & 20.132 & 39.472 & 2.935 \\ 37.022 & 5.411 & 20.132 & 9.035 & 6.759 & 1435 \\ -105.909 & 27.925 & 39.472 & 6.759 & 186.486 & -3422 \\ 12.452 & 0.622 & 2.935 & 1435 & -3422 & 0.490 \end{bmatrix},$$

$$R = \begin{bmatrix} 1 & 0.339 & 0.476 & 0.594 & -0.374 & 0.858 \\ 0.339 & 1 & 0.631 & 0.424 & 0.482 & 0.209 \\ 0.476 & 0.631 & 1 & 0.65 & 0.281 & 0.407 \\ 0.594 & 0.424 & 0.65 & 1 & 0.165 & 0.682 \\ -0.374 & 0.482 & 0.281 & 0.165 & 1 & -0.36 \\ 0.858 & 0.209 & 0.407 & 0.682 & -0.36 & 1 \end{bmatrix}$$

The corresponding vectors of eigen values of S and R are respectively

$$L_s = \begin{bmatrix} 7.615 \\ 3.363 \\ 0.088 \\ 40.966 \\ 205.815 \\ 492.225 \end{bmatrix} \text{ and } L_r = \begin{bmatrix} 0.141 \\ 0.105 \\ 0.337 \\ 0.492 \\ 1.787 \\ 3.137 \end{bmatrix}$$

Development being a multidimensional phenomenon its assessment cannot be completely ascertained by any one of the indicators. Also when a number of indicators are examined individually or collectively, we do not get a comprehensible picture of the distinguishing feature among the various districts. To study the relative development patterns of various districts, we compress the data to managable dimensions (m = 2) by adopting the following two approaches :

a) Obtain feature vectors for which (*i*) loss of information is minimum (*ii*) new variates may have minimum dependence.

We shall choose A so as to minimize the PEM which is a problem of feature extraction or dimensionality reduction in pattern recognition. We select two largest eigen values and find their eigen vectors. Then A is obtained by taking orthogonal eigen vectors corresponding to the largest two eigen values of covariance matrix S as rows.

Thus
$$A = \begin{bmatrix} 0.927 & 0.055 & 0.223 & 0.077 & -0.286 & 0.027 \\ 0.106 & 0.217 & 0.519 & 0.107 & 0.816 & 0.002 \end{bmatrix}$$

The feature vectors that minimise the loss of information are

$$y_1 = 0.927\,x_1 + 0.055\,x_2 + 0.223\,x_3 + 0.077\,x_4 - 0.286\,x_5 + 0.027\,x_6$$

$$y_2 = 0.106\,x_1 + \ + 0.21\,x_2 + 0.519\,x_3 + 0.107\,x_4 + 0.816\,x_5 + 0.002\,x_6$$

On the same lines we obtain A by taking orthogonal eigen vectors corresponding to the largest two eigen values of correlation matrix R as rows.

Thus
$$A = \begin{bmatrix} 0.474 & 0.352 & 0.448 & 0.486 & 0.015 & 0.463 \\ -0.322 & 0.446 & 0.275 & 0.059 & 0.7 & -0.359 \end{bmatrix}$$

Feature vectors that have minimum generalized dependence are

$$z_1 = .474\,x_1 + .352\,x_2 + .448\,x_3 + .486\,x_4 + 0.15\,x_5 + .463\,x_6$$

$$z_2 = -.322\,x_1 + .446\,x_2 + .275\,x_3 + .059\,x_4 + 0.7\,x_5 - 0.359\,x_6$$

Percentage mean squares error in (*i*) and (*ii*) are 6.94 and 17.92 respectively. This indicates that the approach based on variance covariance matrix is better in terms of mean square error criterion. But the second approach is preferred and more suitable when the indicators are measured in different units.

Districts Scores on y_1, y_2, z_1 and z_2 are calculated as given in Table II; y_1 and y_2 have been plotted in Fig 1 (*a*). Similarly Districts Scores on z_1, and z_2 have been plotted in Fig 1 (*b*).

TABLE II

District	y_1	y_2	z_1	z_2
Ambala	83.43	46.25	72.27	6.6
Panchkula	42.25	30.23	41.41	9.63
Yamunanagar	82.9	63.54	78.33	18.99
Kurukshetra	105.57	58.45	93.74	7.07
Kaithal	107.97	57.85	95.29	5.33
Karnal	106.3	51.27	91.17	1.35
Panipat	104.88	49.43	87.02	0.8
Sonipat	96.6	47.47	81.56	3.54
Rohtak	72.09	63.16	72.54	24.57
Faridabad	78.24	40.95	67.1	5.59
Gurgaon	54.29	55.29	60.94	27.01
Rewari	58.39	72.56	67.45	38.31
Mahendergarh	44.87	73.29	61.78	45.1
Bhiwani	47.6	63.75	62.86	35.82
Jind	96.19	68.88	89.68	18.8
Hisar	92.86	82.46	93.4	30.58
Sirsa	80.76	87.44	89.84	39.19

b) Assuming $X = (x_1, x_2, \ldots x_6)$ to be a normal random vector, we can choose two of the six indicators which can be done in 15 different ways. We estimate the entropy $H(X)$ and entropies $H(Y_i)$, i = 1, 2, ..., 15, where $H(Y_i)$ is the entropy of ith combination given in Table III. The loss of information $H(X) - H(Y_i)$ has been evaluated for each selection in Table III. Then we select the pair for which the loss is minimum. In this case the choice of x_1 and x_5 indicators gives minimum loss of information. Districts scores on x_1 and x_5 are plotted in Fig. 2(*a*).

If we consider $\boldsymbol{R}$, correlation matrix of indicators, then

$$\boldsymbol{D} = -\frac{1}{2} \ln |\boldsymbol{R}| \qquad \ldots (6.5)$$

We calculate $\boldsymbol{D}$ for fifteen choices as given in Table III and select that choice for which $\boldsymbol{D}$ is minimum. Thus we get the choice x_4 and x_5 indicators which gives the least measure of dependence D. District Scores on x_4 and x_5 have been plotted in Fig. 2(*b*).

TABLE III

Sr. No. of combinations	Variable in the subset	$\boldsymbol{D}$	Information Loss = $H(X) - H(Y)$
1	x_1, x_2	0.061	9.284
2	x_1, x_3	0.128	8.465
3	x_1, x_4	0.218	9.786
4	x_1, x_5	0.075	8.13 (minimum)
5	x_1, x_6	0.665	11.691
6	x_2, x_3	0.254	10.176
7	x_2, x_4	0.099	11.254
8	x_2, x_5	0.132	9.773
9	x_2, x_6	0.022	12.634
10	x_3, x_4	0.275	10.542
11	x_3, x_5	0.041	8.795
12	x_3, x_6	0.09	11.815
13	x_4, x_5	0.014 (minimum)	10.0
14	x_4, x_6	0.312	13.269
15	x_5, x_6	0.069	11.512

CONCLUDING REMARKS

1. The feature vectors y_1 and y_2 (minimizing loss of information) have higher weights for x_1 and x_5 and thus are pulled towards the indicators having large variances. It is therefore, not surprising to get x_1 and x_5 as the original indicators which also minimize the loss of information.

2. The feature vectors z_1 and z_2 (minimizing the generalized dependence) are obtained by x_4 and x_5 respectively and hence the selection as the least dependent pair is justified. However, the order of 15 pairs obtained by ranking $\boldsymbol{D}$ values is also given by ranking 15 correlations (ignoring signs) between (x_i, x_j) in the correlation matrix s.

3. Figs. 1(*a*) and *b*) it can be observed that there is good resemblance in the various districts in two dimensional space.

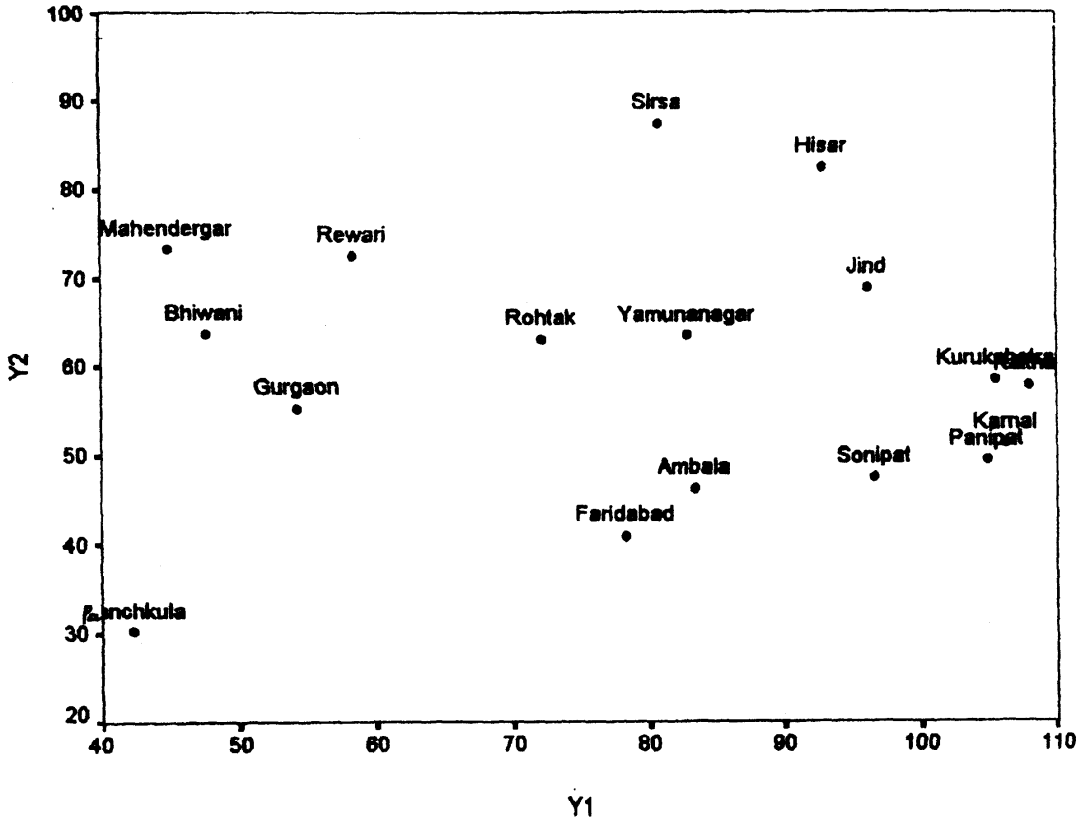
Y2
Y1
Sirsa
Hisar
Mahendergar
Rewari
Jind
Bhiwani
Rohtak
Yamunanagar
Gurgaon
Karnal
Panipat
Sonipat
Ambala
Faridabad

FIG. 1(a)

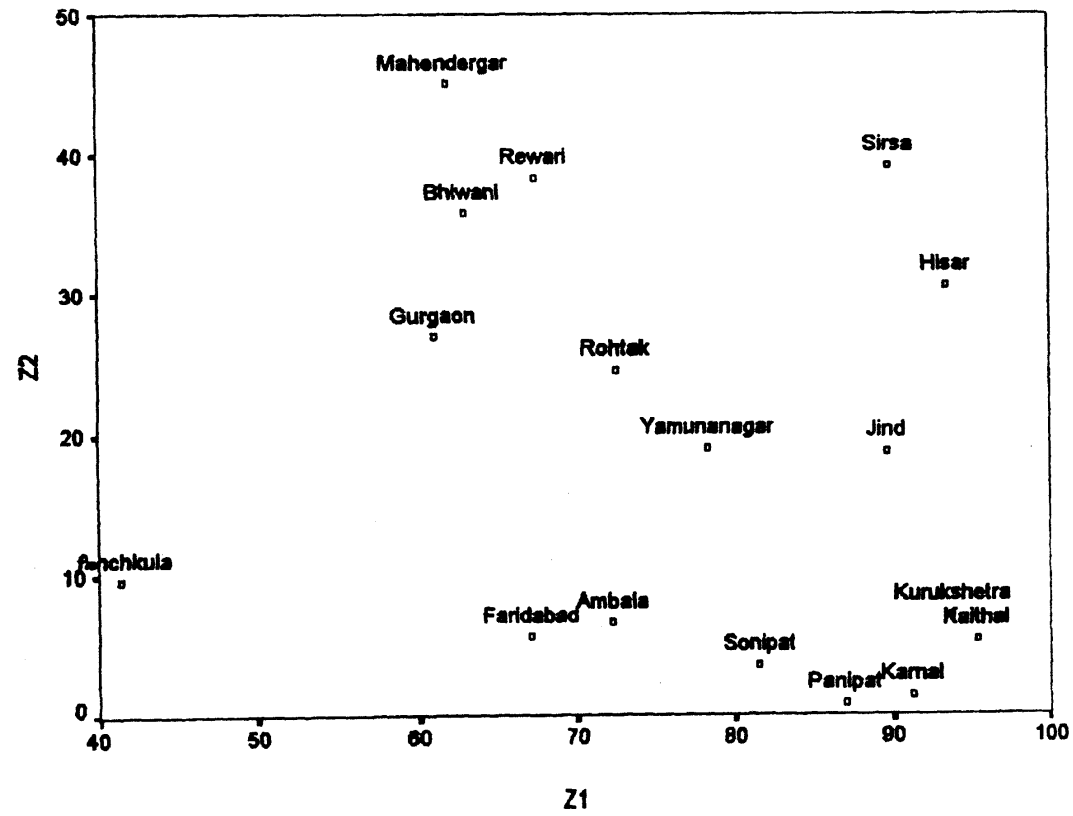
Z2
Z1
Mahendergar
Rewari
Sirsa
Bhiwani
Hisar
Gurgaon
Rohtak
Yamunanagar
Jind
Kurukshetra
Kaithal
Ambala
Faridabad
Sonipat
Karnal
Panipat

FIG. 1(b)

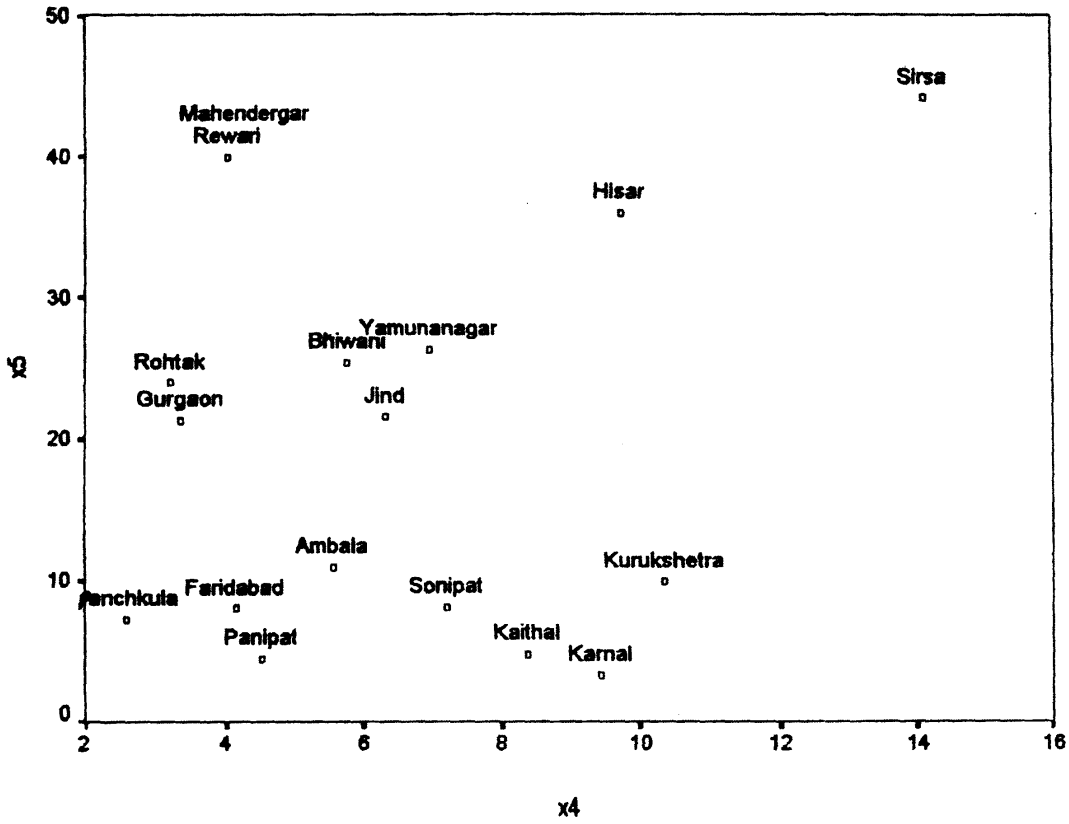
50
40
30
20
10
0
2
4
6
8
10
12
14
16
x4
Sirsa
Mahendergar
Rewari
Hisar
Yamunanagar
Bhiwani
Rohtak
Gurgaon
Jind
Ambala
Kurukshetra
Panchkula
Faridabad
Sonipat
Panipat
Kaithal
Karnal

FIG. 2(a)

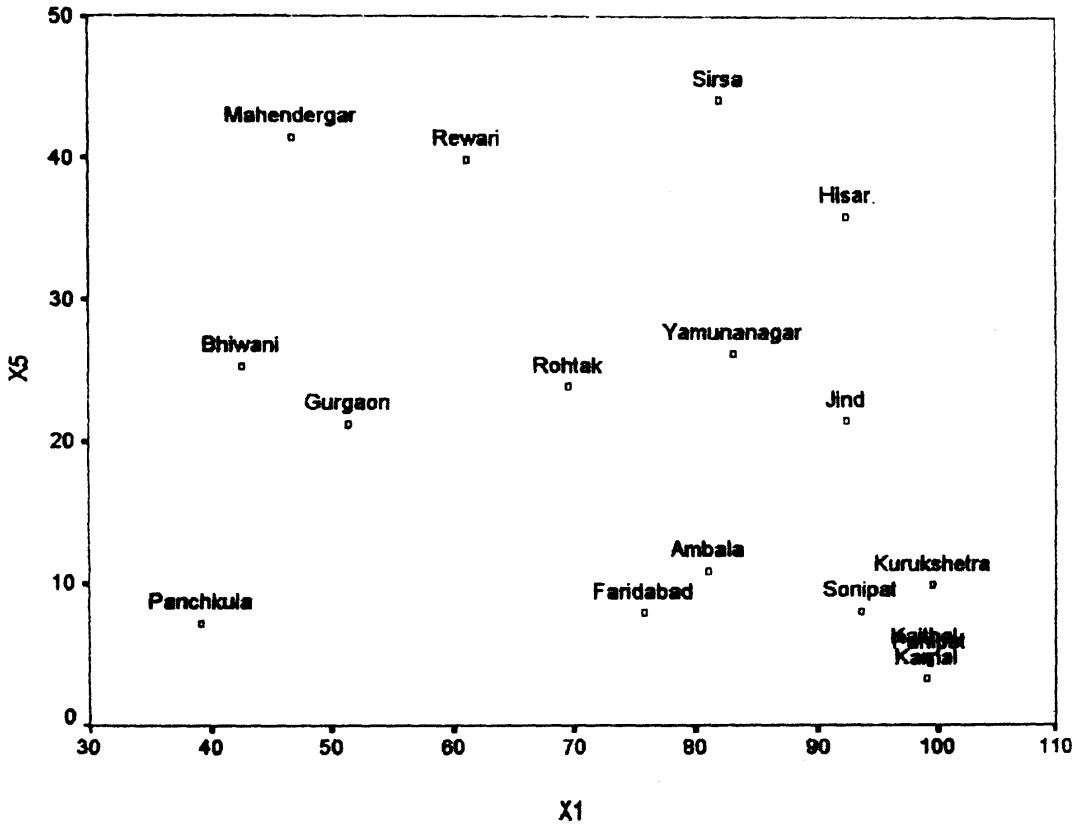
50
40
30
20
10
0
30
40
50
60
70
80
90
100
110
X1
X5
Sirsa
Mahendergar
Rewari
Hisar
Bhiwani
Yamunanagar
Rohtak
Gurgaon
Jind
Ambala
Kurukshetra
Faridabad
Sonipat
Panchkula
Karnal

FIG. 2(b)

4. A comparison of Figs. 2(*a*) and Fig. 2(*b*) shows that pattern of various districts in two dimensional plot is almost similar when we select two principal indicators corresponding to least measure of dependence or minimum loss of Information.

REFERENCES

1. H. Bozdogan. *Commun. Statist. Theor. Method* **19** (1) (1990) 221-78.
2. C. M. Camargo and M. Israel. *Ann. Real Soc. Espan. fits Y qiun* **52A** (1956) 117.
3. C. J. Harris, In: *Recent Theoretical Developments in Control (Ed*: M. J. Gregson) Academic Press, London, 1978.
4. J. N. Kapur, *New Measure of Stochastic Dependence.* IIT/Kanour Res. Rep. No. 243, 1985*a.*
5. J. N. Kapur, *Normalised Measures of Stochastic Dependence.* IIT/Kanpur Res. Rep. No. 279, 1985 *b*
6. J. N. Kapur, *Pattern Recognition* **19** (1986) No. 6, 473-76.
7. J. N. Kapur and M. Dhande, *Acta Ciendica* **XVI M** 2, (1990) 193-98.
8. J. N. Kapur and H. J. Kesavan, *Entropy Optimization Principles with Applications*, Academic Press, New York, 1992.
9. Hea-Jung Kim, *J. Korean statist Soc.* **26** (1) (1997) 131-46.
10. S. Kullback, *Information Theory and Statistics*, Dover Publications, New York, 1968.
11. C. Rajski, *Trans. Third Prague Conf. Inf. Th. Stat. Dec. Funcs. Random Proc.* (1961) 583-85.
12. E. S. Soofi, *J. Amer. statist. Assoc.* **89** (1994) 1243-54.
13. H. Theil and D. G. Fiebig, *Exploiting Continuity; Maximum Entropy Estimation of Continuous Distributions,* Bellinger Publishing Company, Cambridge, Massachusetts, 1984.
14. S. Watanabe, *Knowing and Guessing*, John Wiley, new York, 1969.
15. S. Watanabe, *Pattern Recognition : Human and Mechanical.* John Wiley and Sons, New York, 1985.