Jour. Ind. Soc. Ag. Statistics
51(2&3), 1998 : 379-384

Classification of Observations Using Distances in Oblique Axes System

M.N. Das[1]

SUMMARY

An alternative method of classification based on squares of distances between points corresponding to observation vectors using oblique co-ordinate system has been discussed.

Key words: Observation vectors, p-variate populations, Weight factors, Homogeneous equations, Distance squares.

1. Introduction

The problem of classification in multivariate situation is to devise a rule satisfying certain optimum conditions to place a given observation vector in one of two or more populations from each of which a sample of observation vectors is available. Given two p-variate populations each with variables $(x_1, x_2, \ldots, x_p)$, Fisher [1] proposed $\Sigma\, l_i\, x_i$ as a Discriminant function when l_i's are so obtained as to maximise

$$(\Sigma\, l_i\, x_i)^2 / \Sigma\, l_i\, l_k\, S_{ik}$$

where S_{ik} is the pooled co-variance of the i-th and k-th variables obtained from the samples of the two populations, S_{ii} being the pooled variance of the i-th variable $(i = 1, 2, \ldots, p)$.

No physical meaning can in general be associated with the linear function $L = \Sigma\, l_i x_i$ or its square used in Fisher's method. Hence, when $L^2/\operatorname{var}(L)$ is maximised to obtain l's, it is not clear what exactly is being maximised and why such maximisation need lead to any optimal discriminator. The function, however, has the advantage that it is mathematically tractable for its use to evolve tests provided the assumptions of p-variate normality with equal dispersion matrices of the populations hold as Rao has shown in a series of papers (Rao [2], [3], [4]). These assumptions restrict very much applicability of the function in real life problems. As such it is desirable to evolve procedures having wider scope of applicability even sacrificing on idealisation.

[1] I-1703, C.R. Park, New Delhi-110019

There is also a method known as Bayes procedure which gives a rule for placing a given observation vector into one of two populations such that the cost of misclassification is minimum. This method has, however, the difficulty of determining the cost of misclassification which involves subjectivity and depends on the purpose of investigation and the participating variables.

In the present paper we use an alternative method of classification based on squares of distances between points corresponding to observation vectors using oblique co-ordinate system. As such it has a direct appeal as a discriminator for classification. The method actually uses quadratic forms which are the squares of distances between points in oblique system. Actually, such quadratic forms are natural extension of what may be called distance square, that is, $(x - m)^2$ in univariate case. Fisher's discriminant function is really not linear as a function of observations. It is also a quadratic form in x_i's as $\Sigma\, l_i\, x_i$ becomes actually so when the values of l_i's as functions of x_i's are obtained as indicated earlier and substituted in $\Sigma\, l_i\, x_i$.

2. *Method*

Let $(x_{11}, x_{12}, \ldots x_{1p})$ and $(x_{21}, x_{22}, \ldots x_{2p})$ denote two p-variate observation vectors and $d_i = x_{1i} - x_{2i}$ $(i = 1, 2, \ldots p)$. Treating these two observation vectors as the co-ordinates of two points in a p-dimensional space with oblique axes where w_{ik} is the angle between the i-th and k-th axes $(i = k = 1, 2, \ldots p)$, the square of the distance between the two points is given by $D^2 = \sum\limits_i d_i^2 + 2\, \Sigma\Sigma d_i\, d_k \cos(w_{ik})$.

As the position of a point corresponding to an observation vector depends on the scale of measurements and the weights to be associated with the individual variables for some purpose we can associate with the variables relative weight values, say, l_i's to have a generalised presentation of the co-ordinates of points. Thus, linking, in general, of an observation vector to a point in geometrical spaces is more appropriately provided by $(l_1\, x_1, l_2\, x_2, \ldots, l_p\, x_p)$ so that when an observation vector remains the same its position in space may differ depending on the l-values associated with them. Accordingly, in general,

$$D^2 = \Sigma l_i^2\, d_i^2 + 2\, \Sigma\Sigma l_i\, l_k\, d_i\, d_k \cos(w_{ik})$$

The weight factors also called compounding values by Rao can be chosen suitably or these can also be obtained so as to satisfy certain optimising conditions as will be done here.

Though the angles, w_{ik} have wide choices we propose to link them with the data under investigation by taking cos $(w_{ik}) = r_{ik}$, the correlation coefficient between the i-th and k-th variables.

While dealing with correlated variables this way of linking the angles to data is more appropriate. For example, if for some i and k, $r_{ik} = 1$, the corresponding $w_{ik} = 0$. This means one dimension gets reduced as it should be. This type of dimension reduction is not possible when working with other axes system. Using such linking,

$$D^2 = \Sigma l_i^2 d_i^2 + 2\,\Sigma\Sigma l_i\, l_k\, d_i\, d_k\, r_{ik}$$

2.1 Mean Distance Squares from a Sample of Observation Vectors

Let $(x_{i1}, x_{i2}, \dots x_{ip})$ denote the i-th of n observation vectors from a sample of p-variate observations and $(x_{.1}, x_{.2}, ..., x_{.p})$ denote the vector of means of the variates. Let further $d_{ik} = x_{ik} - x_{.k}$, $(i = 1, 2, \dots n; k = 1, 2, \dots p)$.

Let D_{ik}^2 denote the square of the distance between the points corresponding to the i-th and k-th observation vectors in the sample $(i = k = 1, 2, \dots n)$. It can be shown easily that the average of all such distance squares corresponding to all such possible pairs of the n observation vectors comes out as

$$D_w^2 = \Sigma l_i^2\, S_i^2 + 2\,\Sigma\Sigma\, l_i\, l_k\, S_{ik}\, r_{ik}$$

where S_i^2 is the mean squares of the i-th variate and S_{ik} is the mean S.P. of the i-th and k-th variates $(i = k = 1, 2, \dots p)$.

This is also equal to the average of the n-1 independent distance squares between the points of each observation vector and the mean vector point. Here, r_{ik} is obtained from $S_{ik}/(S_i\, S_k)$.

2.2 Between Samples Distance Squares for Two Samples

Let $(x_{i1}, x_{i2}, \dots x_{ip})$, $(i = 1, 2, \dots, n_1)$ and $(y_{k1}, y_{k2}, \dots y_{kp})$, $(k = 1, 2, \dots, n_2)$ be the observation vectors from two samples of sizes n_1 and n_2 from two p-variate populations. Let further $(x_{.1}, x_{.2}, ..., x_{.p})$ and $(y_{.1}, y_{.2}, ..., y_{.p})$ be the mean vectors for the two samples and $p_i = x_{.i} - y_{.i}$, $(i = 1, 2, \dots p)$.

The square of distance between the points corresponding to the two mean vectors is

$$D_b^2 = \Sigma l_i^2\, p_i^2 + 2\,\Sigma\Sigma l_i\, p_i\, l_k\, p_k\, r_{ik}$$

The within sample pooled mean distance squares is

$$D_w^2 = \Sigma l_i^2 R_i^2 + 2\,\Sigma\Sigma l_i\, l_k\, R_{ik}\, l_{ik}$$

where R_i^2 is the pooled M.S. of the i-th variate over the two samples and R_{ik} is the pooled S.P. of the i-th and k-th variates from the two samples and r_{ik} is the correlation coefficient obtained from the pooled M.S. and pooled mean S.P. of the i-th and k-th variates. That is,

$$r_{ik} = R_{ik}/R_i\, R_k$$

2.3 Estimation of Weight Factors

Using D_b^2 and D_w^2 we propose to use the function

$$L = D_b^2/D_w^2$$

to obtain a discriminator when the weight factors l_i's in it are so obtained as to make the ratio L maximum. Estimates of weight factors are functions of variances and co-variances of the individual variates.

Differentiating log(L) partially with respect to l_i's and equating the differential to zero we get

$$1/L = A/B = k_0, \text{ a constant} \tag{3.1}$$

where $$A = l_i\, p_i^2 + \Sigma_k\, l_k\, p_i\, p_k\, r_{ik} \text{ and } B = l_i\, R_i^2 + \Sigma_k\, l_k\, R_{ik}\, r_{ik}$$

The equation at (3.1) can be written as

$$l_i\,(p_i^2 - k_0\, R_i^2) + \Sigma\, l_k\,(p_i\, p_k - k_0\, R_{ik})\, r_{ik} = 0 \tag{3.2}$$

$$(i = 1, 2, \ldots p)$$

or $$Pl = k_0 Rl \quad (l' = l_1\, l_2\, \ldots l_p)$$

As these are p homogeneous equations we can get non-trivial solutions of l_i's by taking k_0 as a solution of the following determinantal equations in k_0:

$$|\,P - k_0\, R\,| = 0 \text{ where the matrices P and R are as below:} \tag{3.3}$$

$$P = \begin{pmatrix} p_1^2 & p_1 p_2 r_{12} \cdots p_1 p_p r_{1p} \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ p_1 p_p r_{1p} & p_2 p_p r_{2p} \cdots p_p^2 \end{pmatrix}$$

and

$$R = \begin{pmatrix} R_1^2 & R_{12} r_{12} \ldots R_{1p} r_{1p} \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ R_{1p} r_{1p} & R_{2p} r_{2p} \cdots R_p^2 \end{pmatrix}$$

Equation (3.2) can now be written as

$$(P - k_0 R)\, l = 0 \text{ where vector } l' = (l_1, l_2, \ldots, l_p) \tag{3.4}$$

Equations (3.4) can also be written as

$$(PR^{-1} - k_0 I)\, l = 0 \tag{3.5}$$

or

$$(PR^{-1} - k_0 I)\, l = 0 \tag{3.6}$$

From (3.5) it is seen that the solutions for k_0 are the eigen values of PR^{-1}. As $k_0 = 1/L$ the divergence between D_b^2 and D_w^2 is maximum for the minimum eigen value of PR^{-1}. This minimum eigen value is to be used for k_0 in (3.6). To get unique solutions for l_i's the restriction $\Sigma l_i^2 = 1$ is taken.

2.4 Classification

To place an observation vector $(z_1, z_2, \ldots z_p)$ into one of the two populations the following are obtained

$$L_m = \frac{\Sigma_i l_i^2 q_{im}^2 + 2\,\Sigma_i \Sigma_k l_i l_k q_{im} q_{km} r_{ik}}{D_w^2}$$

where $m = 1, 2,$ $q_{i1} = z_i - x_{.1}$ and $q_{i2} = z_i - y_{.1}$

and l_i's are the solutions of (3.6) $(i = 1, 2, \ldots p)$

If $L_1 < L_2$ then the conclusion is that the observation vector under test belongs to the population with mean vector $(x_{.1}, x_{.2}, \ldots x_{.p})$, otherwise it belongs to the other population. In other words the observation vector belongs to that population whose mean vector is nearer to the observation vector in distance.

The functions L_1, L_2 and l_i's are evidently free from units of measurement of the variate values.

3. Classification into One of Several Populations

This technique gets immediately generalised to the case of, say, N populations. In this case D_w^2 is the pooled average distance squares within the samples based on pooled M.S., R_i^2 of the i-th variate and R_{ik}, pooled mean S.P. of the i-th and k-th variates. D_b^2 is similarly the average distance squares of all possible pairs of the mean vector points of the N samples. Actually,

$$D_b^2 = \Sigma l_i^2 P_i^2 + 2\,\Sigma\Sigma l_i\, l_k\, P_{ik}\, r_{ik}$$

where P_i^2 is the mean squares of the sample means of the i-th variate and P_{ik} is the mean S.P. of the means of the i-th and k-th variates.

Taking $L = D_b^2/D_w^2$ and maximising it with respect to l_i's we get as earlier the solutions for l_i's.

Next, the distance square of a point of a given observation vector $(z_1, z_2, \ldots, z_p)$ from that of each of the points of the N mean vectors is obtained. The observation vector belongs to that population whose mean vector point has the least distance from the observation vector point.

4. Discussion

Fisher also maximised $L = D_b^2/D_w^2$ to get l_i's except that in both denominator and numerator of L each r_{ik} has been taken as unity as he used rectangular axes system. As such concept of distance among the points does not hold appropriately in that method.

REFERENCES

[1] Fisher, R.A., 1936. Use of multiple measurements in taxonomic problems. *Ann. Eugen.*, London, 7, 179.

[2] Rao, C.R., 1946. Tests with discriminant functions in multivariate analysis. *Sankhya*, 7, 407.

[3] Rao, C.R., 1948. The utilisation of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc.*, **B10**, 159.

[4] Rao, C.R., 1950. Statistical problems applied to classificatory problems. *Sankhya*, **10**, 229.