THE USE OF DISCRIMINATORY ANALYSIS IN ANTHROPOLOGY

PART I: A SURVEY OF DISCRIMINATORY METHODS

SOŇA DRDKOVÁ

Medical Faculty, Prague — Department of Public Health

Attemps have been made in recent years to generalize the univariate analysis in the case of multiple variates. In multivariate analysis we are concerned with a set of n individuals each of which bears the value of p different variates. The multivariate character lies in the multiplicity of the p variates in the size of the set n. The variates are dependent upon each other so that one or more cannot split off from the others. The variates must be considered together.

The matematical model on which this analysis is based is a multivariate normal distribution.

Classification problems are one of the important questions of multivariate analysis. They led in anthropology and biology to the construction of the coefficient of racial likeness and to the construction of the linear discriminant function.

In 1926 Karl Pearson proposed a measure of racial likeness which has been used since that time by anthropologists for purposes of classifying skeletal remains.

If $n_i^{(1)}$ and $n_i^{(2)}$ denote the number of observations on which the means $\bar{x}_i^{(1)}$ and $\bar{x}_i^{(2)}$ of the i character for the first and second group are based, s_i the standard deviation of the i character and p is the number of characters used, then the coefficient of racial likeness is defined

$$C^2 = \frac{1}{p} \sum_{i=1}^{p} \frac{n_i^{(1)} \cdot n_i^{(2)}}{n_i^{(1)} + n_i^{(2)}} \left(\frac{\bar{x}_i^{(1)} - \bar{x}_i^{(2)}}{s_i} \right)^2$$

This is meant to be an estimate of a measure of distance between two populations. Pearson's paper is in fact the first work on the discriminatory analysis.

Suppose that we have two p-variate populations of a similar kind which "overlap" in the sense that certain members can be observed who might have arisen from either population. There will be such situations that confronting with a member and noting the different values we are uncertain from which population it emanated. We require to assign him to one or to the other of the parents. What rule should we follow in order to make as few mistakes as possible? Questions of this type rise in discriminatory analysis the general object of which is to find rules with optimal properties according to which we can assign individuals to predetermined classes. Note that the classes are predetermined and that we shall in general consider the assignment of a member to one population or to the other. When we assign a member to one of the two populations A or B we can make two kinds of mistakes according to the population to which we have wrongly assigned a given member. We shall suppose that these two kinds of mistakes are equally important. We shall also assume that the two kinds of mistakes occur with equal proportional frequency for each population, that should be as small as possible. We wish to set up a discriminatory boundary.

Suppose that the two populations are multivariate normal with means $\bar{x}_i^{(1)}$ and $x_i^{(2)}$ and identical dispersion matrices. The discriminatory boundary is given by the expression

$$\sum_{i=1}^{p} (a^{1i} d_1 + \ldots + a^{pi} dp)\, x_i = \text{konst.}$$

where

$$d_j = \bar{x}_j^{(1)} - \bar{x}_j^{(2)}$$

This is the linear discriminant function which was first introduced by R. A. Fisher in 1936.

Fisher suggested the following computational procedure. If $x_1, x_2 \ldots x_p$ are measurements then an arbitrary linear compound is

$$l_1 x_1 + l_2 x_2 + \ldots + l_p x_p$$

The coefficients $l_1 \ldots l$ may be chosen such that the linear compound affords the maximum discrimination between the two groups. That is to say we maximize the expression

$$\frac{\{\sum_{j=1}^{p} l_j (x_j^{(1)} - x_j^{(2)})\}^2}{\sum_{ij} l_i l_j a_{ij}}$$

Differentiating with respect to l_i gives us the equations

$$l_1 a_{11} + l_2 a_{12} + \ldots + l_p a_{1p} = d_1$$

$$l_1 a_{21} + l_2 a_{22} + \ldots + l_p a_{2p} = d_2$$

$$l_1 a_{p1} + l_2 a_{p2} + \ldots + l_p a_{pp} = d_p$$

where a_{ij} are the co-variances ($i, j = 1, 2 \ldots p$). By solving the above equations we obtain the coefficients l_i.

Another method of the discriminatory analysis is called the method of distance between two populations.

One important requirement for obtaining anthropologic measurements is to study the possibilities of classifying different groups of individuals in the form of a significant pattern.

The configuration of several groups or of the group characteristics may admit a description in terms of a few group constellations and their interrelationships. The groups within a constellation must necessarily be closer to one another than those belonging to different constellations. The first step in solving that problem of group constellations is the construction of an index by means of which we can measure the resemblance between two groups and compare the distances between any two pairs of those groups. We can say that groups A and B resemble each other more than grous B and C and so on.

Statistical criteria were developed for specifying an individual as a member of one of two groups to which he may possibly belong.

Errors are inevitable in such a procedure and the chances of incorrect classification of individuals of the first and second groups are calculable. It is possible that the two groups overlap to the extent of $100\,\alpha\,\%$. The overlapping is maximal when the groups are identical. It decreases with the increasing of the divergence between the groups. If the groups are distinct in the same sense that the ranges of measurements are non-overlapping, then the percentage of the overlapping iz zero. The extent of separation or divergence between two groups can be judged by α. One might choose a decreasing function of α so that the zero value of α may correspond to the maximum distance. One such function is $1\text{-}\alpha$. This function satisfies the two fundamental postulates of distance:

1. The distance between the groups is not less than zero;

2. The sum of the distance of a group from two other groups is not less than the distance between two other groups.

The distance function must satisfy empirical requirements if it is to be of any value in biological classification:

a) the distance must not decrease when additional characters are considered;

b) the increase in distance by the addition of some characters to a suitable chosen set must be relatively small so that group constellations that were formed on the basis of the chosen set are not distorted when additional characters are considered.

The first requirement is reasonable since adding some characters to a basic set must necessarily reduce the errors of classification. The second requirement has been introduced as a practical necessity. There must be some limit to the number of characters used in order to attain to stable judgements.

The measure $1\text{-}\alpha$ may be replaced by the quantity D^2 which was first introduced by P. C. Mahalanobis. Mahalanobis "Generalized distance" is applicable only to groups in which the measurements are normally distributed. It is given by the expression

$$p\,D_p^2 \sum_i = \sum_j (a^{ij})\,d_i\,d_j$$

where p — the number of characters,

d — mean differences (i = 1, 2 . . . p),

a — co-variances (i, j = 1, 2 . . . p).

The formula for the computation requires the use of computers if $p > 4$.

SUMMARY

This paper contains a popular explanation of some ideas of discriminatory analysis. The other communications which will follow this one, will contain the practical applications of discriminatory analysis in anthropology and some computational schedules.

REFERENCES

T. W. ANDERSON: An Introduction to Multivariate Statistical Analysis. *New York — John Wiley and Sons, 1958.*

R. A. FISHER: The coefficient of racial likeness. *J. Roy. Anthrop. Inst. 66, 57, 1936 b.*

R. A. FISHER: The use of multiple measurement in taxonominic problems. *Ann. Eugen. 1936.*

P. HORST—SMITH STEVENSON: The discrimination of two racial samples. *Psychometrika, 15, 1950.*

P. C. MAHALANOBIS: On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India 12, 1936.*

K. PEARSON: On the coefficients of racial likeness. *Biometrika 18, 1926.*

C. R. RAO: The utilization of multiple measurements in problems of biological classification. *J. Roy. Stat. Soc. B. 10, 1948 a.*