METRIC CONSIDERATIONS IN CLUSTER ANALYSIS

HERMAN CHERNOFF
STANFORD UNIVERSITY

1. Introduction and summary

A variation of the "k means" method of cluster analysis is described which is designed to take into account and profit from the possibility that the separate clusters resemble samples from multivariate normal distributions with substantially different covariance structures. This is preceded by a brief description of a standard version of the method. Indications are given when metric considerations can play an important role and a suitably modified version of the standard method is presented.

While the new method has not yet been applied it is anticipated that its most useful applications will be to situations where the clusters tend to be concentrated in nonparallel hyperplanes of the space of observations. The dimensionality of this space should not be very large. The method should require substantial sample sizes to make the implicit estimates of the covariance matrices useful.

One may expect metric considerations also to be useful in modifying other cluster analysis techniques.

2. The standard k means method

In this section we describe the k means method in the spirit of MacQueen [2]. Suppose that p represents the probability distribution of a random variable (r.v.) Z in an r dimensional Euclidean space and $|y - z|$ represents the distance between points y and z of this space. Let $S = (S_1, S_2, \cdots, S_k)$ be a *decomposition* of the space into k pairwise disjoint measurable subsets (classes) and let $x = (x_1, x_2, \cdots, x_k)$ represent k *reference* points in the space. Then

$$R(x, S) = \sum_{i=1}^{k} \int_{S_i} |z - x_i|^2 \, dp(z) \tag{2.1}$$

is a measure of the corresponding within class variance. From one point of view of the notion of cluster it would be expected that if the probability measure p corresponds to k *natural clusters*, these clusters would relate in a simple way to an (x, S) which minimizes R.

For given S, $R(x, S)$ can be minimized by selecting the reference points to be the centers of gravity, that is, $x = u(S) = (u_1(S), u_2(S), \cdots, u_k(S))$, where