POPULATION MIXTURE MODELS AND CLUSTERING ALGORITHMS

BY

STANLEY L. SCLOVE

TECHNICAL REPORT NO. 11
FEBRUARY 1, 1973

PREPARED UNDER CONTRACT
N00014-67-A-0112-0030 (NR-042-034)
FOR THE OFFICE OF NAVAL RESEARCH

THEODORE W. ANDERSON, PROJECT DIRECTOR

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

POPULATION MIXTURE MODELS AND CLUSTERING ALGORITHMS*

BY

STANLEY L. SCLOVE
University of Illinois at Chicago Circle

TECHNICAL REPORT NO. 11

FEBRUARY 1, 1973

PREPARED UNDER THE AUSPICES

OF

OFFICE OF NAVAL RESEARCH CONTRACT # N00014-67-A-0112-0030

THEODORE W. ANDERSON, PROJECT DIRECTOR

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

*Also issued as Technical Report No. 71, National Science Foundation Grant GP-32326X.

POPULATION MIXTURE MODELS AND CLUSTERING ALGORITHMS

by

Stanley L. Sclove
University of Illinois at Chicago Circle

Abstract.

The problem of clustering individuals is considered within the context of a mixture of distributions. A modification of the usual approach to population mixtures is employed. As usual, a parametric family of distributions is considered, a set of parameter values being associated with each population. In addition, with each observation is associated an identification parameter, indicating from which population the observation arose. The resulting likelihood function is interpreted in terms of the conditional probability density of a sample from a mixture of populations, given the identification parameter of each observation. Clustering algorithms are obtained by applying a method of iterated maximum likelihood to this likelihood function.

AMS 1970 subject classification. 62H30; Secondary 62E10.

Key words and phrases. Mixture of distributions, cluster analysis, isodata procedure, k-means procedure, Mahalanobis distance.

OUTLINE

POPULATION MIXTURE MODELS AND CLUSTERING ALGORITHMS

by

Stanley L. Sclove
University of Illinois at Chicago Circle

Summary.

The problem of clustering individuals is considered within the context of a mixture of distributions. A modification of the usual approach to population mixtures is employed. As usual, a parametric family of distributions is considered, a set of parameter values being associated with each population. In addition, with each observation is associated a parameter indicating from which population the observation arose. The resulting likelihood function is interpreted as the conditional probability density of a sample from the mixture of populations, given the population identifications of each observation.

The relation of this conditional mixture model to the standard mixture model is discussed; it is shown how the concept of the conditional mixture model provides a probability model for cluster analysis, and it is shown how to use the model to provide a plausible general method for clustering.

Given a parametric family of distributions, an appropriate clustering algorithm is obtained by applying a method of iterated maximum likelihood to the resulting likelihood function. The algorithms resulting by application of this general method are, then, interpretable as schemes for estimating the parameters of probability models.

Special attention is given to the case of multivariate normal populations with common covariance matrix. This case is of special interest because application of the general method produces Mahalanobis-distance versions of two well-known clustering algorithms, isodata and k-means, thereby relating these algorithms to a probability model for the clustering problem. Other models given special attention are the multivariate normal distribution with different covariance matrices, and multinomial models, especially the model based on an assumption of local independence as used in latent structure analysis.

1. Introduction.

The problem of "clustering" to be considered here is as follows: given a sample of p-vectors $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$, that is, a sample of p observations on each of n individuals, put the individuals into groups. Of course the problem needs more formalization if we are to be able to do anything meaningful with it.

We begin by defining a clustering as a partition of the set of observations, that is, a collection $\{C_1, C_2, \ldots, C_k\}$ of disjoint sets such that each observation belongs to one and only one set C_g. Each set C_g $(g=1,\ldots,k)$ is a cluster.

In this paper we shall assume that the integer k is specified in advance. (A modification of the algorithm to be presented allows some of the clusters to join or split, thereby permitting fewer or more than k clusters to be formed. See Section 6.2.)

As has been suggested before [see, e.g., Fleiss and Zubin (1969)], it seems reasonable to consider a population mixture model for clustering problems. With the g-th population is associated the probability density function $h_g(\underset{\sim}{x})$, g=1,...,k. When we are working with some parametric family, say indexed by a parameter $\underset{\sim}{\beta}$, h_g takes the form $h_g(\underset{\sim}{x}) = h(\underset{\sim}{x}; \underset{\sim}{\beta}_g)$. The densities (or parameters) are unknown, this being the distinction between the present formulation of the clustering problem and the classical classification problem, sometimes termed "identification", "discrimination", or "allocation". In the classical problem, the densities or parameters are known, or else a training set of data is available, from which the densities or parameters can be estimated.

Now with Individual i (i=1,...,n) associate the group identification parameter γ_i which is equal to g if and only if Individual i belongs to group g(g=1,2,...,k). Each individual gives rise to a pair $(\underset{\sim}{X},\gamma)$. $\underset{\sim}{X}$ is observable; γ is not. It will thus be seen that this problem fits into the framework of an empirical Bayes problem [see, e.g., Robbins (1964)], but in the present paper this aspect will not be studied explicitly.

In the terminology used by Neyman and Scott (1948) in a study of consistent estimation, the parameters γ_i are "incidental" parameters because each of them refers to a finite number of observations (one in the present case), while the parameters $\underset{\sim}{\beta}_g$ are "structural" parameters because, if we allow n to tend to infinity, each of them is associated with an infinite number of observations.

In the context of this model, to "cluster" is merely to estimate the γ_i's, i=1,...,n individuals.

It is convenient to reparametrize somewhat. Replace γ_i by the k-vector $\underset{\sim}{\theta}_i$ which consists of k-1 zeros and a single 1, the position of the 1 indicating which group Individual i belongs to; that is, $\underset{\sim}{\theta}_i$ has a 1 as its γ_i-th element and 0's elsewhere. The density of $\underset{\sim}{X}_i$, given θ_i, is

$$f(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) = \sum_{g=1}^{k} \theta_{gi} h_g(\underset{\sim}{x}_i) \quad , \tag{1.1}$$

where θ_{gi} is the g-th component of $\underset{\sim}{\theta}_i$.

2. The probability model.

This model should be compared and contrasted with the usual population mixture model, in which any observation $\underset{\sim}{x}_i$ is chosen from Population g with probability π_g, so that the density of $\underset{\sim}{X}_i$ is

$$j(\underset{\sim}{x}_i; \pi_1,\ldots,\pi_k) = \sum_{g=1}^{k} \pi_g h_g(\underset{\sim}{x}_i) \quad . \tag{2.1}$$

The probability model that will be used here for the clustering problem is as follows. It is assumed that pairs $(\underset{\sim}{X}_i, \underset{\sim}{\Theta}_i)$, i=1,...,n, have been sampled randomly, in the sense that their joint density is

$$\prod_{i=1}^{n} f_{\underset{\sim}{X}_i, \underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i, \underset{\sim}{\theta}_i) \quad . \tag{2.2}$$

(The notation which will be used here is the standard notation in which f and F are generic symbols for probability density functions and cumulative distribution functions, respectively, f_X denotes the probability density function of the random variable X, $f_{X,Y}$ denotes the joint density of X and Y, $f_{Y|X}$ denotes the conditional probability density function of Y, given X, etc. For the moment we suppress the subscript i.)

The conditional density of $\underset{\sim}{X}$ given $\underset{\sim}{\Theta}$ is

$$f_{\underset{\sim}{X}|\underset{\sim}{\Theta}}(\underset{\sim}{x}|\underset{\sim}{\theta}) = \sum_{g=1}^{k} \theta_g h_g(\underset{\sim}{x}) \quad .$$

The marginal density of $\underset{\sim}{\Theta}$ is taken to be the point multinomial,

$$f_{\underset{\sim}{\Theta}}(\underset{\sim}{\theta}) = \pi_1^{\theta_1} \pi_2^{\theta_2} \cdots \pi_k^{\theta_k} \quad ,$$

$\theta_g = 0$ or 1, $\Sigma_{g=1}^{k} \theta_g = 1$, $\pi_g > 0$, $\Sigma_{g=1}^{k} \pi_g = 1$. Thus π_g is the probability that a randomly selected individual comes from Population g.

First it will be shown that the standard mixture density is indeed the marginal density for $\underset{\sim}{X}$ resulting from this model. Somewhat more generally, <u>let $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_k)$ be a random vector. If the conditional density of $\underset{\sim}{X}$ given $\underset{\sim}{Z}$ is</u>

$$f_{\underset{\sim}{X}|\underset{\sim}{Z}}(\underset{\sim}{x}|\underset{\sim}{z}) = \sum_{g=1}^{k} z_g h_g(\underset{\sim}{x}) \quad ,$$

then the marginal density of $\underset{\sim}{X}$ is

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}) = \sum_{g=1}^{k} E[Z_g] h_g(\underset{\sim}{x}) \quad .$$

To see this, note that we have

$$\begin{aligned} f_{\underset{\sim}{X}}(\underset{\sim}{x}) &= \int f_{\underset{\sim}{X},\underset{\sim}{Z}}(\underset{\sim}{x},\underset{\sim}{z})\, d\underset{\sim}{z} \\ &= \int f_{\underset{\sim}{X}|\underset{\sim}{Z}}(\underset{\sim}{x}|\underset{\sim}{z})\, f_{\underset{\sim}{Z}}(\underset{\sim}{z})\, d\underset{\sim}{z} \\ &= \int \sum_{g=1}^{k} z_g h_g(\underset{\sim}{x})\, f_{\underset{\sim}{Z}}(\underset{\sim}{z})\, d\underset{\sim}{z} \\ &= \sum_{g=1}^{k} [\int z_g f_{\underset{\sim}{Z}}(\underset{\sim}{z}) d\underset{\sim}{z}]\, h_g(\underset{\sim}{x}) \\ &= \sum_{g=1}^{k} E[Z_g]\, h_g(\underset{\sim}{x}) \quad . \end{aligned}$$

From this it follows that <u>if</u> $\underset{\sim}{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_k)$ <u>has the point multinomial density</u>

$$f_{\underset{\sim}{\Theta}}(\underset{\sim}{\theta}) = \pi_1^{\theta_1} \pi_2^{\theta_2} \cdots \pi_k^{\theta_k} \quad ,$$

$\pi_g > 0$, $\sum_{g=1}^{k} \pi_g = 1$, $\theta_g = 0$ or 1, $\sum_{g=1}^{k} \theta_g = 1$, <u>then the marginal density of</u> $\underset{\sim}{X}$ <u>is the standard mixture density</u>:

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}) = \sum_{g=1}^{k} \pi_g h_g(\underset{\sim}{x}) \quad .$$

For, under this model the random variable $Z_g = \Theta_g$ is Bernoulli with parameter π_g; hence $E[\Theta_g] = \pi_g$.

Now suppose that pairs $(\underset{\sim}{X}_i, \underset{\sim}{\Theta}_i)$, i=1,...,n, are sampled randomly, in the sense that their joint density is (2.2). Of course, then the $\underset{\sim}{X}$'s are independent, and the $\underset{\sim}{\Theta}$'s are independent. Then the conditional density of $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$, given $\underset{\sim}{\Theta}_1, \underset{\sim}{\Theta}_2, \ldots, \underset{\sim}{\Theta}_n$ is

$$(2.3)\quad f_{\underset{\sim}{X}_1,\ldots,\underset{\sim}{X}_n|\underset{\sim}{\Theta}_1,\ldots,\underset{\sim}{\Theta}_n}(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n|\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n)$$

$$= \frac{f_{\underset{\sim}{X}_1,\underset{\sim}{\Theta}_1,\ldots,\underset{\sim}{X}_n,\underset{\sim}{\Theta}_n}(\underset{\sim}{x}_1,\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{x}_n,\underset{\sim}{\theta}_n)}{f_{\underset{\sim}{\Theta}_1,\ldots,\underset{\sim}{\Theta}_n}(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n)}$$

$$= \frac{\prod_{i=1}^{n} f_{\underset{\sim}{X}_i,\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i,\underset{\sim}{\theta}_i)}{\prod_{i=1}^{n} f_{\underset{\sim}{\Theta}_i}(\underset{\sim}{\theta}_i)}$$

$$= \frac{\prod_{i=1}^{n} f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) f_{\underset{\sim}{\Theta}_i}(\underset{\sim}{\theta}_i)}{\prod_{i=1}^{n} f_{\underset{\sim}{\Theta}_i}(\underset{\sim}{\theta}_i)}$$

$$= \prod_{i=1}^{n} f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) \quad .$$

It is (2.3) which is the "likelihood" in the conditional population mixture model. In the context of this model, then, to "cluster" is to estimate the $\underset{\sim}{\theta}_i$'s, the values of the identification parameters.

Versions of this model have been used recently by Scott and Symons (1971) and S. John (1970), but the model dates back at least to Gibson (1959), where it is called the latent profile model. This model has been discussed by Anderson (1959).

The likelihood approach to clustering is illuminating in that it sometimes shows how ad hoc optimality criteria (objective functions) which have been proposed for the clutering problem relate to particular probability models. For example, Scott and Symons (1971) show how various optimality criteria relate to maximum likelihood clustering in multivariate normal populations.

Note that we can equivalently write (1.1) as a product:

$$f(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) = \prod_{g=1}^{k} [h_g(\underset{\sim}{x}_i)]^{\theta_{gi}} . \tag{2.4}$$

The form (2.4) is often more convenient, and we shall use it in what follows.

It is easy to allow for the presence of a "training set" of data -- a prior set of observations for each of which we know the group identification. Letting m_g be the number of prior observations in the g-th group and denoting the prior observations from the g-th group by $\underset{\sim}{w}_{g\ell}$, $\ell=1,\ldots,m_g$, we can write the likelihood as

$$\prod_{g=1}^{k} \prod_{\ell=1}^{m_g} h_g(\underset{\sim}{w}_{g\ell}) \prod_{i=1}^{n} \prod_{g=1}^{k} [h_g(\underset{\sim}{x}_i)]^{\theta_{gi}} ,$$

if we treat all the observations $\underset{\sim}{W}_{g\ell}$, g=1,...,k, ℓ=1,...,m_g, and $\underset{\sim}{X}_i$, i=1,...,n as statistically independent. We do not explicitly treat the case of prior observations any further here.

3. The clustering algorithm.

Using the form (2.4), one sees that under the random sampling mechanism mentioned above the joint probability density function of $\underset{\sim}{X}_1,\underset{\sim}{X}_2,\ldots,\underset{\sim}{X}_n$, given $\underset{\sim}{\Theta}_1,\underset{\sim}{\Theta}_2,\ldots,\underset{\sim}{\Theta}_n$ is

$$\prod_{i=1}^{n}\prod_{g=1}^{k}[h_g(\underset{\sim}{x}_i)]^{\theta_{gi}} ,$$

or, in parametric form,

$$\prod_{i=1}^{n}\prod_{g=1}^{k}[h(\underset{\sim}{x}_i;\underset{\sim}{\beta}_g)]^{\theta_{gi}} .$$

The likelihood is to be maximized over all assignments of individuals to groups and over all permissible parameter values. Many ad hoc schemes can be applied to this maximization problem. For example, one way to maximize is to start with a given clustering $C_1,\ldots,C_k$, take each observation successively and shift it to the first cluster for which a shift results in an increase in likelihood, and loop through the data until no individual changes clusters.

The algorithm to be described here is an iterated, that is, a back-and-forth procedure of maximizing this likelihood function, in that we first maximize with respect to the θ's (holding the $\underset{\sim}{\beta}$'s fixed at initial

values), then we maximize with respect to the $\underset{\sim}{\beta}$'s (holding the θ's fixed at the values obtained in the previous stage), then we again maximize with respect to the θ's (holding the $\underset{\sim}{\beta}$'s fixed at the values obtained in the previous stage), etc. We stop when no θ changes, i.e., when no individual changes clusters -- or when we have used a pre-specified amount of computer time.

An alternative for starting the procedure is to start with an initial clustering rather than with initial guesses of the $\underset{\sim}{\beta}$'s.

It is clear that, for fixed values of the $\underset{\sim}{\beta}$'s, say $\hat{\underset{\sim}{\beta}}$'s, the likelihood is maximized, for each i, by taking

$$\hat{\theta}_{gi} = 1 \quad \text{if} \quad h(\underset{\sim}{x}_i;\underset{\sim}{\beta}_g) = \max_{1\le\ell\le k} \{h(\underset{\sim}{x}_i;\hat{\underset{\sim}{\beta}}_\ell)\}$$

$$= 0 \quad \text{otherwise.} \tag{3.1}$$

(In case of ties an arbitrary choice is made.) In other words, clustering proceeds by allocating Individual i to that group for which the estimated probability density of the observation $\underset{\sim}{x}_i$ is largest.

Note that, having tentatively estimated the γ~s (or, equivalently, the θ's) at any stage, that is, having tentatively clustered the individuals, estimation of the $\underset{\sim}{\beta}$'s is reduced simply to ordinary maximum likelihood estimation in the particular parametric family at hand.

Let T denote the set of $\underset{\sim}{\theta}_i$'s and B the set of $\underset{\sim}{\beta}_g$'s. Write $L(B,T)$ to denote the likelihood. Let $B^{(s)}$ denote the value of B

which maximizes L at the s-th stage of the iteration, and similarly let $T^{(s)}$ denote the value of T which maximizes L at the s-th stage of the iteration. Then $T^{(s)}$ maximizes $L(B^{(s)},T)$ with respect to T, and $B^{(s)}$ maximizes $L(B,T^{(s-1)}$ with respect to B. As a function of B, $L(B,T^{(s-1)})$ is the section of $L(B,T)$ at $T=T^{(s-1)}$ and $L(B^{(s)},T)$ as a function of T is the section of $L(B,T)$ at $B=B^{(s)}$. We may refer to this back-and-forth maximization as section-wise maximization. It is an example of the relaxation method (or "Southwell's method"); see Ortega and Rheinboldt (1970, pp. 214ff.) and Southwell (1940 and 1946).

It is true that

$$L(B^{(s+1)},T^{(s)}) \geq L(B^{(s)},T^{(s)}) \quad ,$$

and

$$L(B^{(s)},T^{(s+1)}) \geq L(B^{(s)},T^{(s)})$$

that is, at no stage of the procedure can the value of the likelihood be decreased; however, there is no guarantee of convergence to the global maximum (neither do alternative clustering algorithms guarantee convergence to the global maximum of their objective functions).

To see how the procedure can fail to converge to a global maximum, suppose it happens that $L(B^{(s)},T^{(s)}) > L(B,T^{(s)})$, for all B, or $L(B^{(s)},T^{(s-1)}) > L(B^{(s)},T)$, for all T. Then the procedure will terminate at the s-th stage, without having necessarily reached a global maximum.

That is, if, having maximized with respect to one of the variables B and T, we happen to find ourselves at a (relative) maximum with respect to the other, we may not reach a global maximum.

Back-and-forth iterative methods such as the one developed here are familiar in other estimation problems, notably in weighted least squares estimation, where we iterate between estimating the weights and the regression coefficients, and in factor analysis, where we iterate between estimating the communalities and the factor loadings.

4. Application to particular distributions.

Now we consider application of this general clustering method to particular families of distributions. First we consider normal distributions with common covariance matrix, for it is in this case that it becomes clear how the model establishes a link with some existing clustering procedures.

4.1. Multivariate normal populations with common covariance matrix.

In the case of p-variate normal populations with means $\underset{\sim}{\mu}_g$, $g=1,\ldots,k$, and common covariance matrix $\underline{\Sigma}$, the likelihood takes this form:

$$(2\pi)^{-np/2} \, |\underline{\Sigma}|^{-n/2} \exp\left[-\tfrac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{k} \theta_{gi} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)' \, \underline{\Sigma}^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)\right] .$$

Here (3.1) is equivalent to

$$\hat{\theta}_{gi} = 1 \quad \text{if} \quad (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_g)' \hat{\underline{\Sigma}}^{-1} (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_g) = \min_{1 \le \ell \le k} [(\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_\ell)' \hat{\underline{\Sigma}}^{-1} (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_\ell)]$$

$$= 0 \quad \text{otherwise.} \tag{4.1}$$

That is, Individual i is assigned to that group to whose tentatively-estimated centroid he is closest, where the distance is in the metric of the tentatively estimated covariance matrix. Having estaimted the θ's, we have multivariate normal observations arranged into groups; maximization with respect to the $\underset{\sim}{\mu}$'s and $\underline{\Sigma}$ is accomplished by taking the sample mean vectors as estimates for the $\underset{\sim}{\mu}$'s, and the within-groups sum-of-products matrix gives the estimate of $\underline{\Sigma}$. The procedure is iterated: using new estimates $\hat{\underset{\sim}{\mu}}_g$, $g=1,\ldots,k$, and $\hat{\underline{\Sigma}}$, (4.1) is applied again. Then new $\hat{\underset{\sim}{\mu}}$'s and a new $\hat{\underline{\Sigma}}$ are calculated, etc. The matrix $\hat{\underline{\Sigma}}$ can be updated efficiently. Also, the Mahalanobis distances in (4.1) can be efficiently computed as follows. These distances are of the form $\underset{\sim}{v}'\underline{M}^{-1}\underset{\sim}{v}$, where $\underset{\sim}{v} = (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_g)$ and $\underline{M} = \hat{\underline{\Sigma}}$. To evaluate a quadratic form $\underset{\sim}{v}'\underline{M}^{-1}\underset{\sim}{v}$, given $\underline{M}$ and $\underset{\sim}{v}$, one notes that, algebraically, the solution $\underset{\sim}{x}$ of the system $\underline{M}\underset{\sim}{x} = \underset{\sim}{v}$ is $\underset{\sim}{x} = \underline{M}^{-1}\underset{\sim}{v}$. Numerically, this solution $\underset{\sim}{x}$ can be obtained efficiently, without doing all the arithmetic operations required to obtain $\underline{M}^{-1}$. One then computes the value of $\underset{\sim}{v}'\underline{M}^{-1}\underset{\sim}{v}$ simply as $\underset{\sim}{v}'\underset{\sim}{x}$. [See Anderson (1958), p. 107.]

Relationship with the "isodata" procedure. This scheme is a Mahalanobis-distance version of Ball and Hall's (1967) isodata clustering procedure. (Earlier documentation of isodata by Ball and Hall exists, but the 1967 reference is perhaps the most accessible.) The isodata scheme

proceeds as follows. One starts with tentative estimates of cluster means and assigns each individual to the mean to which he is closest. (The isodata scheme uses Euclidean distance, or modified Euclidean distance in which different weights are assigned to the p dimensions.) The cluster means are then re-estimated, and one loops through the data again, reassigning the individuals, etc. Note the similarity to our scheme. We start with tentative estimates of the $\underset{\sim}{\mu}$'s and $\underline{\Sigma}$ (it seems a good idea to take the initial estimates of the $\underset{\sim}{\mu}$'s to be outside the convex hull of the data, and it is easy to take the initial estimate of $\underline{\Sigma}$ to be the identity matrix), and assign each individual to the mean to which he is closest, using Mahalanobis-distance in the metric of the tentatively estimated covariance matrix. The $\underset{\sim}{\mu}$'s and $\underline{\Sigma}$ are then re-estimated, the individuals are re-allocated to clusters, etc.

An important difference is that our scheme employs Mahalanobis-distance rather than Euclidean or weighted-Euclidean distance. And it is worth emphasizing that it is the Mahalanobis distance based on the within-groups sum-of-products matrix that arises here; some data analysts use the total sum-of-products matrix, which, as Chernoff (1970), for example, has argued, is not appropriate. I have done data analyses using both the total and the within-groups sum-of-products matrices, and the total sum-of-products matrix gave poor results, while the within-groups sum-of-products matrix gave good results.

For example, consider the Fisher iris data [Fisher (1936)], consisting of $p=4$ measurements on each of 50 irises in each of $k=3$ species. If the sample centroids of the three species are computed from the group-identified data and the 150 flowers are then assigned to that centroid to which they are "closest", then only three misclassifications are made when the distance is in the metric of the within-groups covariance matrix, 11 misclassifications are made if Euclidean distance is used, and 20 misclassifications are made when the distance is in the metric of the total covariance matrix!

One further point along these lines: Mahalanobis-distance is the same as Euclidean distance in terms of principal axes. Hence some data analysts transform the raw data into scores on principal components, so that they can simply use Euclidean distance. Their mistake is that they use the principal components of the <u>total</u> sum-of-products matrix. The Euclidean distance they calculate is then the same as Mahalanobis-distance in the metric of the <u>total</u> sum-of-products matrix, which is not appropriate.

I have programmed three algorithms in APL [I.B.M. (1969), Iverson (1962)] -- the algorithm developed here, in which at any stage the distance is in the metric of the tentatively-estimated covariance matrix, an algorithm in which Euclidean distance is used at each stage, and an algorithm in which at each stage the distance was in the metric of the total covariance matrix. Results of two rums of each of the three algorithms on the Fisher iris data will be given here. In one run, the initial centroids (initial estimates of the three mean vectors) were flowers in the <u>same</u> species ("difficult initial centroids"). In another

run, the initial centroids were three flowers from the three different species.

RESULTS OF TWO RUNS OF EACH OF THREE ALGORITHMS

Metric	Difficult initial centroids		Easy initial centroids	
	Number of misclassifications	Number of iterations before convergence	Number of misclassifications	Number of iterations before convergence
The adaptive metric of the algorithm, starting with $\hat{\Sigma}=I$ (Euclidean distance)	6	14	3	5
Euclidean distance	16	11	16	3
Distance in the metric of the total sum-of-products matrix	40	10	29	6

Relationship with the "k-means" procedure. Arranging the computation a little differently, updating the estimates of the $\underset{\sim}{\mu}$'s and Σ after each individual is assigned rather than waiting until all individuals have been assigned, produces a Mahalanobis-distance version of MacQueen's (1966) k-means procedure.

Thus, a link has been established between some of the better known ad hoc clustering procedures and a probability model for the clustering problem.

4.2. Multivariate normal populations with different covariance matrices.

The algorithm generated for this case turns out not to be simply to use a different Mahalanobis distance for each cluster. The complication which occurs is analogous to that in "classical" classification (discriminant analysis), where one is led to quadratic discriminant functions if the covariance matrices differ.

The details are as follows. The likelihood in this case is

$$(2\pi)^{-np/2} \prod_{i=1}^{n} \prod_{g=1}^{k} |\underset{\sim}{\Sigma}_g|^{-\theta_{gi}/2} \exp\Big[-\tfrac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{k} \theta_{gi} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)' \underset{\sim}{\Sigma}_g^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}_g)\Big] .$$

In this case (3.1) becomes

(4.2) $\hat{\theta}_{gi} = 1$ if setting $\ell = g$ maximizes

$$|\hat{\underset{\sim}{\Sigma}}_\ell|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_\ell)' \hat{\underset{\sim}{\Sigma}}_\ell^{-1} (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_\ell)]$$

$= 0$ otherwise.

Maximizing the expression in (4.2) is equivalent to minimizing

$$\ell n |\hat{\underset{\sim}{\Sigma}}_\ell| + (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_\ell)' \hat{\underset{\sim}{\Sigma}}_\ell^{-1} (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_\ell) .$$

It has been noted [see, e.g., Day (1969)] that in the standard mixture model for this case the supremum of the likelihood is infinity. This is reflected in the fact that in our algorithm it would be possible that at some stage one of the clusters would consist of a single individual, so that the tentative estimate of the mean of that cluster would be the

vector of observations for that individual, and the tentative estimate of the covariance matrix of that cluster would be undefined. It is also possible for the observations in a given cluster to be very close to lying on a lower-dimensional subspace, so that the tentative estimate of the covariance matrix could have an arbitrarily small determinant, and the maximized likelihood could be arbitrarily large, for the contribution of Cluster g to the maximized likelihood is $|\hat{\Sigma}_g|^{-ng/2} \exp(-pn_g/2)$, where n_g is the number of individuals assigned to that cluster.

4.3. Multinomial models.

Multinomial models are of special interest because they relate to the analysis of questionnaires and of patterns of medical symptoms. Suppose each variable X_v, $v=1,\ldots,p$, is a dichotomous variable (indicating a Yes or No answer, or presence or absence of a symptom). It is not reasonable to assume the X's independent in the whole (mixture) population. It is sometimes assumed, however, that within subpopulations, they are independent. This model of local independence is employed in latent structure analysis [Lazarsfeld and Henry (1968)]. If we let

$$\beta_{vg} = \Pr\{X_{vi}=1\} = 1 - \Pr\{X_{vi}=0\} ,$$

for the g-th subpopulation, then under the assumption of local independence the density in the g-th subpopulation is

$$h_g(\underset{\sim}{x}_i;\underset{\sim}{\beta}_g) = \prod_{v=1}^{p} \beta_{vg}^{x_{vi}}(1-\beta_{vg})^{1-x_{vi}} .$$

The "clusters" are the subpopulations.

5. Comparison with the method based on the standard mixture model.

Wolfe (1970) has considered clustering based on the standard mixture model. Under that model, the posterior probability that Individual i belongs to Group g is

$$\frac{\pi_g h_g(\underset{\sim}{x}_i;\underset{\sim}{\beta}_g)}{\sum_{\ell=1}^{k} \pi_\ell h_\ell(\underset{\sim}{x}_i;\underset{\sim}{\beta}_\ell)} \quad . \tag{5.1}$$

If we can obtain estimates for $\underset{\sim}{\beta}_g$, π_g, $g=1,\ldots,k$, they can be substituted to provide an estimate of (5.1),

$$\frac{\hat{\pi}_g h(\underset{\sim}{x}_i;\hat{\underset{\sim}{\beta}}_g)}{\sum_{\ell=1}^{k} \hat{\pi}_\ell h(\underset{\sim}{x}_i;\hat{\underset{\sim}{\beta}}_\ell)} \quad . \tag{5.2}$$

Individual i is assigned to that Group g for which the estimated posterior probability of group membership (5.2) is largest. (Recall that, with the conditional mixture model, Individual i is assigned to that Group g for which the estimated density $h(\underset{\sim}{x}_i;\hat{\underset{\sim}{\beta}}_g)$ is largest.)

Wolfe has provided computer programs for the case of normal distributions. As is well known, the maximum likelihood equations for mixture problems are messy. Wolfe solves them by a multivariate Newton-Raphson method. This involves the assignment of arbitrary initial values

to the parameters, to start the iterative solution, as does the general method described here.

Perhaps a word may be said by way of further comparison of the standard and conditional mixture models. The likelihood in the standard model is

$$\prod_{i=1}^{n} f_{\underset{\sim}{X}_i}(\underset{\sim}{x}_i) \quad ,$$

or

$$\prod_{i=1}^{n} \int f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i)\, dF_{\underset{\sim}{\Theta}_i}(\underset{\sim}{\theta}_i) \quad ,$$

whereas the likelihood in the conditional model is

$$\prod_{i=1}^{n} f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) \quad ,$$

so that in using the conditional model we are using the factors $f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i)$ rather than a smoothed version of them, namely

$$\int f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i)\, dF_{\underset{\sim}{\Theta}_i}(\underset{\sim}{\theta}_i) = E\left[f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i)\right] = f_{\underset{\sim}{X}_i}(\underset{\sim}{x}_i) \quad .$$

Note that

$$\max_{\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n} \prod_{i=1}^{n} f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) = \max_{\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n} \prod_{i=1}^{n}\prod_{g=1}^{k} [h_g(\underset{\sim}{x}_i)]^{\theta_{gi}}$$

$$= \prod_{i=1}^{n} \max_{\underset{\sim}{\theta}_i} \prod_{g=1}^{k} [h_g(\underset{\sim}{x}_i)]^{\theta_{gi}}$$

$$= \prod_{i=1}^{n} \max_{\underset{\sim}{\theta}_i} f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i)$$

$$\geq \prod_{i=1}^{n} \int f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i)\, dF_{\underset{\sim}{\Theta}}(\underset{\sim}{\theta})$$

$$= \prod_{i=1}^{n} f_{\underset{\sim}{X}_i}(\underset{\sim}{x}_i)$$

$$= \prod_{i=1}^{n} j(\underset{\sim}{x}_i;\pi_1,\ldots,\pi_k) \quad ,$$

no matter what the values of $\pi_1,\ldots,\pi_k$. Thus

$$\max_{\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n} \prod_{i=1}^{n} f_{\underset{\sim}{X}_i|\underset{\sim}{\Theta}_i}(\underset{\sim}{x}_i|\underset{\sim}{\theta}_i) \geq \max_{\pi_1,\ldots,\pi_k} \prod_{i=1}^{n} j(\underset{\sim}{x}_i;\pi_1,\ldots,\pi_k) \quad ,$$

i.e.,

$$\max_{\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n} L(\underset{\sim}{\theta}_1,\ldots,\underset{\sim}{\theta}_n;\ \underset{\sim}{X}_1,\ldots,\underset{\sim}{X}_n) \geq \max_{\pi_1,\ldots,\pi_k} L'(\pi_1,\ldots,\pi_k;\ \underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n) \quad ,$$

where $L'(\pi_1,\ldots,\pi_k;\ \underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n) = \prod_{i=1}^{n} j(\underset{\sim}{x}_i;\ \pi_1,\ldots,\pi_k)$ denotes the likelihood corresponding to the standard model. If it is legitimate to compare likelihoods under the two different models, this shows how "overfit" occurs when we use conditional models; the same concepts apply to the "shrinkage problem" in regression analysis when we predict using an estimated regression function.

Note that, under the assumption or random sampling from the k populations, the π_g's of the standard model can be estimated after clustering based on the conditional model; we can take as the estimate the proportion of individuals assigned to Population g:

$$\hat{\pi}_g = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{gi} = \frac{n_g}{n} \quad ,$$

where $n_g = \Sigma_{i=1}^{n} \hat{\theta}_{gi}$ is simply the number of individuals assigned to Population g. That is, under an assumption of random sampling, we can use results obtained from working with the conditional distribution of $\underset{\sim}{X}$ to estimate parameters of the marginal distribution of $\underset{\sim}{X}$.

These two types of models, conditional and unconditional, arise in other statistical contexts as well, notably analysis of variance [Eisenhart's (1947) classification of effects as "Model I" or "Model II" is now standard] and factor analysis [see Anderson and Rubin (1956)].

6. Some remarks on statistical inference.

Again let $L(B,T)$ denote the likelihood as a function of the structural parameters B and the incidental parameters T, given the data. The maximum likelihood estimate of (B,T) is the value $(\hat{B},\hat{T})$ for which L is largest. The quantity $L(\hat{B},\hat{T})$ is the corresponding maximum value of the likelihood. To approximate $(\hat{B},\hat{T})$, one uses the algorithm. Let

$$\lambda(B,T) = L(B,T)/L(\hat{B},\hat{T}) \quad .$$

Let F denote the asymptotic (as n tends to infinity) cumulative distribution function of $-2 \ln \lambda(B,T)$: $\lim_{n\to\infty} \Pr\{-2 \ln \lambda(B,T) \leq x\} = F(x)$.

Suppose that F is independent of (B,T). For example, it may be the cumulative distribution function of a chi-square distribution with an appropriate number of degrees of freedom; it is necessary to investigate the extent to which the large sample theory of the generalized likelihood ratio applies when there are incidental parameters.

6.1. Confidence sets.

Let x_α denote the upper-α percentage point of F. Then

$$1 - \alpha = F(x_\alpha) = \Pr\{-2 \ln \lambda(B,T) \leq x_\alpha\} = \Pr\{-2 \ln L(B,T) \leq x_\alpha + 2 \ln L(\hat{B},\hat{T})\} \ ,$$

so that

$$\{(B,T): -2 \ln L(B,T) \leq x_\alpha + 2 \ln(\hat{B},\hat{T})\}$$

is an approximate $100(1-\alpha)\%$ confidence set for (B,T).

Denote by $(\tilde{B},\tilde{T})$ the estimates produced by the algorithm. Then $L(\tilde{B},\tilde{T}) \leq L(\hat{B},\hat{T})$. Thus a conservative confidence set -- one that contains more values of (B,T) than the true confidence set and has confidence coefficient at least $1-\alpha$ -- is

$$\{(B,T): -2 \ln(B,T) \leq x_\alpha + 2 \ln(\tilde{B},\tilde{T})\} \quad .$$

6.2. Some remarks on choice of k.

The algorithm can be run with different choices of k and the results can be compared. Note that the likelihood function is a different function for different values of k. Denote this dependence upon k by denoting the likelihood by $L_k(B_k,T_k)$. Let $\hat{B}_k$, $\hat{T}_k$ denote the maximum likelihood estimates. Following Wolfe's approach for the standard mixture model, one might make a sequence of hypothesis tests to decide on k, first comparing $L_2(\hat{B}_2,\hat{T}_2)$ with $L_3(\hat{B}_3,\hat{T}_3)$, then if necessary comparing $L_3(\hat{B}_3,\hat{T}_3)$ with $L_4(\hat{B}_4,\hat{T}_4)$, etc. Wolfe uses the asymptotic chi-square distribution of the generalized likelihood ratio here; even in the context of the standard mixture model this may not be the asymptotic distribution.

An alternative approach to the choice of k, is to follow a suggestion of MacQueen and introduce refinement and coarsening parameters R and C such that two clusters coalesce when their centroids are less than R units

apart and a cluster splits when its diameter (maximum distance between any two of its members) exceeds C.

7. Conclusions.

A modification of the usual mixture model has been employed to provide a probability framework for clustering problems. A general method of producing clustering algorithms which correspond to a method of iterated maximum likelihood has been given. The general method given here is a plausible method for clustering which is linked to a probability model and which is comparatively easy to program. In the case of multivariate normal distributions with common covariance matrix the general method produces clustering schemes which can be viewed as improved versions of some existing schemes.

The focus here has been on the parametric case, but the methods discussed might be applied to the nonparametric case by estimating the densities $h_g(\underset{\sim}{x})$ as the clustering proceeds, using standard methods of density estimation.

Clustering algorithms based on a likelihood function are based on the raw data matrix, in contradistinction to many clustering procedures which are based on a matrix of pairwise similarities or distances. The latter procedures have the advantage of applicability to problems where a raw data matrix is not available. When the raw data are available, such algorithms have the disadvantage of not extracting all the information from the observations and the computational disadvantage of preliminary computation of all the pairwise distances (or similarities).

Acknowledgements. This research has been supported by National Science Foundation Grants GP-22595 at Carnegie-Mellon University and GP at Stanford University and Office of Naval Research Contract #N00014-67-A-0112-0030 (NR-042-034) at Stanford University.

REFERENCES

[1] Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York.

[2] Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. Probability and Statistics: The Harald Cramer Volume, U. Grenander, ed., 9-38. Almqvist and Wiksell, Uppsala.

[3] Anderson, T. W., and Rubin, Herman. (1956). Statistical inference in factor analysis. Proc. Third Berkeley Symposium Math. Statist. and Prob., J. Neyman, ed., 5, 111-150. University of California Press, Berkeley and Los Angeles.

[4] Ball, G. H., and Hall, David J. (1967). A clustering technique for summarizing multivariate data. Behavioral Sciences 12, 153-155.

[5] Chernoff, Herman. (1970). Metric considerations in cluster analysis. Proc. Sixth Berkeley Symposium Math. Statist. and Prob. 1, 621-629.

[6] Day, N. E. (1969). Estimating the components of a mixture of normal distributions. Biometrika 56, 463-475.

[7] Eisenhart, C. (1947). The assumptions underlying the A.O.V. Biometrics 3, 1-21.

[8] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179-188.

[9] Fleiss, J. L., and Zubin, J. (1969). On the methods and theory of clustering. Multivariate Behavioral Research 4, 235-250.

[10] Gibson, W. A. (1959). Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. Psychometrika 24, 229-252.

[11] International Business Machines Corp. (1969). APL-360 Primer, 2nd ed., (IBM Publication GH20-0689-1). IBM Corporation, Technical Publications Dept., White Plains, New York.

[12] Iverson, Kenneth E. (1962). A Programming Language. John Wiley and Sons, Inc., New York.

[13] John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics 12, 553-563.

[14] Lazarsfeld, Paul F., and Henry, Neil W. (1968). Latent Structure Analysis. Houghton Mifflin Co., Boston.

[15] MacQueen, J. (1966). Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symposium Math. Statist. and Prob. 1, 281-297.

[16] Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations, with particular reference to structural relations. Econometrica 16, 1-32.

[17] Ortega, James, and Rheinboldt, Werner. (1970). Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York.

[18] Robbins, Herbert. (1964). The empirical Bayes approach to statistical decision problems. Ann. Math. Statist. 35, 1-20.

[19] Scott, A. J., and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics 27, 387-397.

[20] Southwell, R. (1940). Relaxation Methods in Engineering Science: A Treatise on Approximate Computation. Oxford University Press, London.

[21] Southwell, R. (1946). Relaxation Methods in Theoretical Physics. Oxford University Press (Clarendon), London and New York.

[22] Wolfe, John H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research 5, 329-350.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY, CALIF.

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE
POPULATION MIXTURE MODELS AND CLUSTERING ALGORITHMS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
TECHNICAL REPORT

5. AUTHOR(S) (Last name, first name, initial)
SCLOVE, Stanley L.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
February 1, 1973	28	22

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0112-0030	# 11
b. PROJECT NO.	
c NR-042-034)	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	#71 NSF GP-32326X

10. AVAILABILITY/LIMITATION NOTICES
Reproduction in whole or in part is permitted for any purpose of the United States Government

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Arlington, Va.

13. ABSTRACT

The problem of clustering individuals is considered within the context of a mixture of distributions. A modification of the usual approach to population mixtures is employed. As usual, a parametric family of distributions is considered, a set of parameter values being associated with each population. In addition, with each observation is associated an identification parameter, indicating from which population the observation arose. The resulting likelihood function is interpreted in terms of the conditional probability density of a sample from a mixture of populations, given the identification parameter of each observation. Clustering algorithms are obtained by applying a method of iterated maximum likelihood to this likelihood function.

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
mixture of distributions						
cluster analysis						
isodata procedure						
k-means procedure						
Mahalanobis distance						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manuel. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.