CLUSTERING CRITERIA AND MULTIVARIATE NORMAL MIXTURES

by

M.J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1258

December 1979

CLUSTERING CRITERIA AND MULTIVARIATE NORMAL MIXTURES

M.J. Symons*

Department of Biostatistics
University of North Carolina
Chapel Hill, N.C. 27514, U.S.A.

SUMMARY

New clustering criteria are presented for use when a mixture of multivariate normal distributions is an appropriate model. They are derived from maximum likelihood and Bayesian approaches, corresponding to different assumptions about the covariance matrices of the mixture components. Two of these are modifications of the determinant of the within groups sum of squares criterion of Friedman and Rubin (1967). These appear to be more sensitive to disparate cluster sizes. Two others are appropriate for different shaped clusters.

The performance of these criteria, and another one studied by Maronna and Jacovkis (1974) for heterogeneous covariance matrices, is compared in an example requiring the separation of two types of diabetic patients from normal subjects. The results with the three criteria appropriate for different shaped clusters were comparable to one another and preferable to those from three criteria for similar shaped clusters.

Key Words & Phrases: Clustering criteria, mixtures of multivarite normals, variable metric.

---

*This research was completed while the author was on leave with an Intergovernmental Personnel Act appointment with the Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, N.C. 27711.

1. INTRODUCTION AND MODEL

Several authors, including Wolfe (1967, 1969), Day (1969), Scott and Symons (1971), and Binder (1978), have directed attention to solving clustering problems with a mixture of multivariate normals as a statistical model. For the number of components in the mixture specified as G, the clustering problem has been formulated as one of estimating the mixture component origin of each of the n p-variate observations, $\underset{\sim}{y}_i$, where i indexes the sample.

More specifically the density of $\underset{\sim}{y}_i$ is

$$f(\underset{\sim}{y}_i | \pi_g\text{'s}, \underset{\sim}{\mu}_g\text{'s}, \Sigma_g\text{'s}) = \sum_{g=1}^{G} \pi_g N_p(\underset{\sim}{y}_i | \underset{\sim}{\mu}_g, \Sigma_g), \qquad (1)$$

where the π_g's are the mixing parameters, each required to be positive and that they sum to unity, and the notation $N_p(\underset{\sim}{y}_i | \underset{\sim}{\mu}_g, \Sigma_g)$ denotes that $\underset{\sim}{y}_i$ is distributed as a p-variate normal with mean vector $\underset{\sim}{\mu}_g$ and covariance matrix Σ_g.

The unknown mixture component origin for $\underset{\sim}{y}_i$ is denoted by z_i. From the mixture model we have that z_i equals g, or equivalently $\underset{\sim}{y}_i$ comes from the g-th component, with probability π_g. The clustering problem then can be viewed as one of estimating the n z_i.

2. CLUSTERING CRITERIA

For both the maximum likelihood and Bayesian approaches, the likelihood of the data is required in order to estimate the n components in the vector $\underset{\sim}{z} = (z_1, z_2, \ldots, z_n)$. Letting the matrix Y denote the n observations $\underset{\sim}{y}_i$ and $\underset{\sim}{\theta}$ be the vector of parameters $(\pi_1, \ldots, \pi_G, \underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_G, \Sigma_1, \ldots, \Sigma_G)$, the likelihood of the data then is given by

$$L(Y | \underset{\sim}{\theta}, \underset{\sim}{z}) = \prod_{g=1}^{G} \left\{ \pi_g^{n_g} |\Sigma_g|^{-\frac{1}{2}n_g} \right\} \exp\left\{ -\tfrac{1}{2} \sum_{g=1}^{G} \sum_{C_g} (\underset{\sim}{y}_i - \underset{\sim}{\mu}_g)' \Sigma_g^{-1} (\underset{\sim}{y}_i - \underset{\sim}{\mu}_g) \right\}, \qquad (2)$$

where C_g is the collection of $\underset{\sim}{y}_i$'s with $z_i = g$, and n_g is the number of observations in C_g. This likelihood is conditional on the parameters $\underset{\sim}{\theta}$ and a specified vector $\underset{\sim}{z}$, allocating each $\underset{\sim}{y}_i$ to one of the G components in the mixture (1). As there are G^n possible allocations, approximate search routines are required to find the optimal assignment of the observations to the groups and thereby to determine the clusters $C_1, C_2, \ldots, C_G$.

The maximum likelihood (ML) approach determines the ML estimate of $\underset{\sim}{z}$, $\hat{\underset{\sim}{z}}$, as the allocation that maximizes (2). The parameters are replaced by their ML estimates given an allocation of the n $\underset{\sim}{y}_i$ to the G components. These ML estimates are the standard ones, namely

$$\hat{\pi}_g = n_g/n \ , \tag{3}$$

$$\hat{\underset{\sim}{\mu}}_g = \bar{\underset{\sim}{y}}_g = \frac{1}{n_g} \sum_{C_g} \underset{\sim}{y}_i , \tag{4}$$

and

$$\hat{\sum}_g = \frac{1}{n_g} W_g = \frac{1}{n_g} \sum_{C_g} (\underset{\sim}{y}_i - \bar{\underset{\sim}{y}}_g)(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}}_g)' , \tag{5}$$

for $g = 1, \ldots, G$.

The Bayesian approach, discussed generally by Lindley (1966), requires a specification of a prior distribution for $\underset{\sim}{\theta}$, $p(\underset{\sim}{\theta})$. Jeffreys' (1961) priors were utilized in order to simply delineate the bounds of the parameter space. As the parameters, $\underset{\sim}{\theta}$, are not of central interest, the product of the likelihood (1) and prior $p(\underset{\sim}{\theta})$ is averaged over. Specifically,

$$L(Y|\underset{\sim}{z}) = \int L(Y|\underset{\sim}{\theta}, \underset{\sim}{z}) p(\underset{\sim}{\theta}) d\underset{\sim}{\theta}, \tag{6}$$

where the integration is over the parameter space of $\underset{\sim}{\theta}$. Geisser (1966) refers to the normalization of (6) as the predictive distribution of Y. The mode of (6) is taken as the Bayes estimate of the optimal allocation, $\tilde{\underset{\sim}{z}}$.

Clustering criteria for these two approaches are presented for the cases when the mixture components in (1) have unknown covariance matrices that are assumed to be homogeneous and when they are not necessarily equal. The component origin of all observations is presumed unknown. None of the parameters in (1) is presumed known. For other cases and situations with some parameters specified and/or with the component origin of some observations known, see Geisser (1966), Scott and Symons (1971), Symons (1973) and Binder (1978). These references also contain more of the details of the Bayesian approach to this problem.

2.1 Covariance Matrices Homogeneous

When $\sum_g = \sum$ for $g = 1, \ldots, G$ and $\sum$ is unknown, the ML approach is to maximize the likelihood over G^n possible allocations. For each allocation z, the ML estimators (3), (4) and

$$\hat{\sum} = \frac{1}{n} W = \frac{1}{n} \sum_{g=1}^{G} W_g \ , \tag{7}$$

replace the parameters $\underset{\sim}{\theta}$. The ML optimal allocation, $\hat{\underset{\sim}{z}}$, is equivalent to the partition of the n observations into G groups which minimizes the criterion

$$n \ell n[|W|] - 2 \sum_{g=1}^{G} n_g \ell n[n_g], \tag{8}$$

where $|W|$ is the determinant of the within groups sum of squares and $\ell n[\ ]$ denotes the natural logarithm.

The Bayesian approach utilizes a vague prior, namely,

$$p(\underset{\sim}{\theta}) = p(\pi_1, \ldots, \pi_G) p(\underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_G | \sum) p(\sum) \propto \left[\prod_{g=1}^{G} \pi_g \right]^{-1} |\sum|^{-\frac{1}{2}(p+1)}, \tag{9}$$

to define the parameter space. The prior on $\sum$ is of the general form used by Geisser and Cornfield (1963). The product of (1) and (9) is

averaged over the parameter space of $\underset{\sim}{\theta}$, as described by (6). The Bayesian optimal allocation, $\tilde{\underset{\sim}{z}}$, is equivalent to the partition of the data into G groups that minimizes the criterion.

$$(n-G)\ell n[|W|] + \sum_{g=1}^{G}\left\{p\ell n[n_g] - 2\ell n[\Gamma(n_g)]\right\}. \tag{10}$$

The ML criterion (8) and Bayes criterion (10) are modifications of the determinant of the within groups sum of squares, a criterion proposed by Friedman and Rubin (1967). It has been observed by Scott and Symons (1971) and Binder (1978) that the $|W|$ criterion tends to favor partitions of equal size. Criteria (8) and (10) can be shown to be more sensitive to disparate group sizes for situations when the shape of the clusters is similar. On the other hand, as can be seen from the example in Section 3, these same two criteria may tend to create different sized but more homogeneous clusters when applied to heterogeneous clustering problems. Criteria (8) and (10) provide very similar cluster results.

2.2 Covariance Matrices Unequal and Unknown

When the covariance matrices may differ from group to group, the ML optimal allocation $\hat{\underset{\sim}{z}}$, is that partition of the n observations into G groups which minimizes the criterion

$$\sum_{g=1}^{G} n_g \ell n[|W_g|] - 2\sum_{g=1}^{G} n_g \ell n[n_g] \ . \tag{11}$$

The difference between criteria (11) and (8) is that the within groups sum of squares matrices, W_g from (3), are required to estimate an elipsoidal shape for each of the G components in the mixture. So that none of the W_g will be singular, each cluster must be required to contain at least $p+1$ observations.

The Bayesian approach for this case utilizes vague prior infromation as in (9), but independently for each of the G $\sum_g$. The optimal allocation, $\tilde{z}$, is determined by (6) and is equivalent to the partition minimizing the criterion

$$\sum_{g=1}^{G}\left\{(n_g - 1)\ell n[|W_g|] + p\ell n[n_g] - p(n_g + p)\ell n[2] - 2\left[\ell n[\Gamma(n_g)] + \sum_{i=1}^{p}\ell n[\Gamma(\tfrac{1}{2}\{n_g + p + 1 - i\})]\right]\right\} . \quad (12)$$

This Bayes criterion is considerably more complicated than the ML one presented in (11). The additional detail comes from the normalization by (6) of a Wishart density for each of the G components in (1). Only with small samples would one expect a difference in the performance of criteria (11) and (12) or (8) and (10).

One additional criterion for different shaped clusters is presented and compared with (11) and (12) in the next Section. Maronna and Jacovkis (1974) compared several multivariate clustering procedures with variable metrics. Their study included numerical experiments using real data and Monte Carlo data simulating mixtures of G normal p-variate populations. The variable metric, based upon the suggestions of Chernoff (1970) and especially Rohlf (1970), was the only one "for which the clustering method had reasonable properties". It was shown to be equivalent to the clustering criterion

$$\sum_{g=1}^{G}|W_g|^{1/p} , \quad (13)$$

and is based upon the within cluster covariance matrix normalized to have unit determinant. By showing this equivalence between a metric clustering procedure and optimization of a clustering criterion, or "uncertainty functional", Maronna and Jacovkis have provided an

important connection between quite different approaches to cluster analysis.

3. EXAMPLE

The relationship between chemical diabetes and overt diabetes in 145 non-obese adult subjects was examined by Reaven and Miller (1977). The degree of glucose intolerance, insulin response to oral glucose, and insulin resistance in normal subjects and patients with non-ketotic diabetes was determined. The three dimensional shape of the data set was that of a "boomerang with two wings and a fat middle", and is reproduced in Figure 1. The two wings were interpreted as representing patients with chemical diabetes and overt diabetes, respectively. The spherical middle corresponded to normal subjects.

The heterogeneity of the parts of this three dimensional shape and their non-ellipsoidal nature provide a practical test for the criteria presented in Section 2. Reaven and Miller (1977) utilized the determinant of the within groups sum of squares as their clustering criterion, supplemented by the means for three groups from an earlier set of 125 patients, that were similar to those groups presented in Figure 1. This a priori information was not used in the analyses reported here, as the capability of these clustering criteria alone was of primary interest.

These criteria were all options in an approximate routine constructed by McRae (1971) with slight modifications. After determining the partition that corresponded to a minimum value of the criterion selected from among 32 randomly generated partitions, the routine produces a relative minimum for the selected criterion in the sense that any re-assignment of one observation results in a larger

value. This algorithm does not guarantee an absolute minimum over the G^n possible allocations, in fact numerous local minima were often found. Because of this several runs were made with variations in the algorithm. Also the partitions in the neighborhood of the clinical classification, based upon the results of the glucose tolerance test and reported by Reaven and Miller (1977), were examined.

Table 1 contains the results of the application of the determinant of the within groups sum of squares, the ML and Bayesian modifications of the same, (8) and (10), respectively, the ML (11) and Bayes (12) criterion for heterogeneous clusters, and Maronna and Jacovkis variable metric equivalent criterion (13). The notation: (73, 10,1)(3, 26,6) (0, 0,26) gives the resulting composition of the three clusters found by Reaven and Miller, relative to the clinical classification given at the head of Table 1. Specifically, the first cluster (73, 10,1) is composed of 73 normals, 10 chemical diabetics and one overt diabetic. The second cluster is composed of three normals, 26 chemical diabetics and six overt diabetics, while 26 overt diabetics form the third cluster.

There are four table entries for each of the six criteria. First is the criterion value for the clinical classification. Second is the clustering with a minimum criterion value found by starting at the clinical classification. The third and fourth entries are the partitions corresponding to the two smallest criterion values located by McRae's (1971) program.

The best solution is unknown for this data set. However, the combined impression from Figure 1, the clinical classification, and the results obtained by Reaven and Miller utilizing the results from 125 earlier patients, suggest that the criteria (11), (12) and (13)

have produced reasonable clusters. These criteria are appropriate for different shaped clusters. The results from the Bayes criterion (12) and the variable metric criterion (13) appear slightly better than those with the ML criterion (11) for this example. The result with $|W|$ is not nearly as good. The results with the ML (8) and Bayes (10) modification of $|W|$ can be very poor. These latter three criteria are appropriate for homonogeneous shaped clusters and their performance can clearly be misleading when different shaped clusters are present.

There seems to be no simple recommendation to guide users of these criteria. The determinant of the within groups sum of squares can be insensitive to different sized clusters. The maximum likelihood (8) and Bayes (10) modifications of $|W|$ may produce very unbalanced clusters when presented with heterogeneous clusters. Some preliminary empirical experience with the criteria (11), (12) and (13), which are appropriate for different shaped clusters, suggests that these criteria may produce clusters of different shapes and sizes when presented with homogeneous shaped clusters that are close together. Improved software may alleviate these difficulties in some instances, but there are still multiple local minima present in many data sets. There appears to be no substitute for careful evaluation of the results obtained from different analyses with several two dimensional scatter plots.

4. FURTHER DISCUSSION

It is worthwhile to note that

$$L(Y|\underset{\sim}{\theta}) = \prod_{i=1} \left[\sum_{g=1}^{G} \pi_g N_p(\underset{\sim}{y}_i | \underset{\sim}{\mu}_g, \Sigma_g) \right] = \sum L(Y|\underset{\sim}{\theta}, \underset{\sim}{z}), \tag{14}$$

where the summation of (1) at the right is over all G^n allocations of the n $\underset{\sim}{y}_i$ to the G components. It is the likelihood (14) that is maximized with respect to $\underset{\sim}{\theta}$ by Wolfe (1967, 1969) and Day (1969). Given the ML estimate of $\underset{\sim}{\theta}$ then each of the n observations $\underset{\sim}{y}_i$ is assigned to the component for which $\hat{\pi}_g N_p(\underset{\sim}{y}_i | \hat{\underset{\sim}{\mu}}_g, \hat{\Sigma}_g)$ is largest. With large samples this procedure is optimal and fortunately relieves the concern for searches over G^n allocations to optimize criteria such as those presented in Section 2. However these same criteria appear to perform better with small samples and can profitably be used to provide initial estimates, (3), (4), and (5) or (7), for a program to maximize (14), as noted by Symons (1973).

Marriott (1975) has pointed out that such estimates of $\underset{\sim}{\theta}$ from a partition $\hat{\underset{\sim}{z}}$ or $\tilde{\underset{\sim}{z}}$ are inconsistent. Clearly the estimates (3), (4), and (5) or (7) are conditional on $\hat{\underset{\sim}{z}}$. Estimates based upon (14) are then unconditional, as regards any partition. That these conditional estimates are inconsistent is intuitively reasonable as one can see that differences between means will be over-estimated and variances under-estimated whenever there is overlap in the mixture components and estimates are computed given such partitions of the data. It should be kept in mind however, that in cluster analysis the interest is on the estimation of the best allocation of the n observations to the G groups, as noted by Scott and Symons (1971), Symons (1973) and Binder (1978). If estimation of the parameters of the model (1) is primarily of interest, then the likelihood in (14) is the cornerstone of theoretically sound estimation of $\underset{\sim}{\theta}$ from a maximum likelihood or a Bayesian perspective.

REFERENCES

Binder, D.A. (1978). Bayesian cluster analysis. *Biometrika 65*, 31-38.

Chernoff, H. (1970). Metric considerations in cluster analysis. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Univ. of Calif. Press, 621-629.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*, 463-74.

Friedman, H.P. and Rubin, J. (1967). On some invariant criterion for grouping data. *J. Amer. Stat. Ass. 63*, 1159-1178.

Geisser, S. (1966). Predictive discrimination. *Multivariate Analysis*. Proceedings of an International Symposium. Edited by P.R. Krishnaiah. Academic Press, New York and London, 149-63.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society, Series B 25*, 368-76.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon, Oxford.

Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*. Cambridge University Press.

Maronna, R. and Jacovkis, P.M. (1974). Multivariate clustering procedures with variable metrics. *Biometrics 30*, 499-505.

Marriott, F.H.C. (1975). Separating mixtures of normal distributions. *Biometrics 31*, 767-769.

McRae, D.J. (1971). MIKCA: a Fortran IV iterative K-means cluster analysis program. *Behavioral Science 16*, 423-4.

Reaver, G.M. and Miller, R. (1977). An inquiry into the nature of diabetes mellitus using a multidimensional analysis. *Technical Report No. 33*, Division of Biostatistics, Stanford University, Stanford, California.

Rohlf, F.J. (1970). Adaptive hierarchical clustering schemes. *Syst. Zool. 19*, 58-82.

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

Symons, M.J. (1973). Bayes modification of some clustering criteria. *Institute of Statistics Mimeo Series No. 880*, Department of Biostatistics, Univ. of North Carolina at Chapel Hill.

Wolfe, J.H. (1967). NORMIX: computation methods for estimating the parameters of multivariate normal mixtures of distributions. *Research Memo, SRM 68-2.* U.S. Naval Personnel Research Activity, San Diego, California.

Wolfe, J.H. (1969). Pattern clustering by multivariate mixture analysis. *Research Memo. SRM 69-17.* U.S. Naval Personnel Research Activity. San Diego, California.

FIGURE 1

Artist's rendition of data as seen in three dimensions
(Reaven and Miller (1977, p. 25))

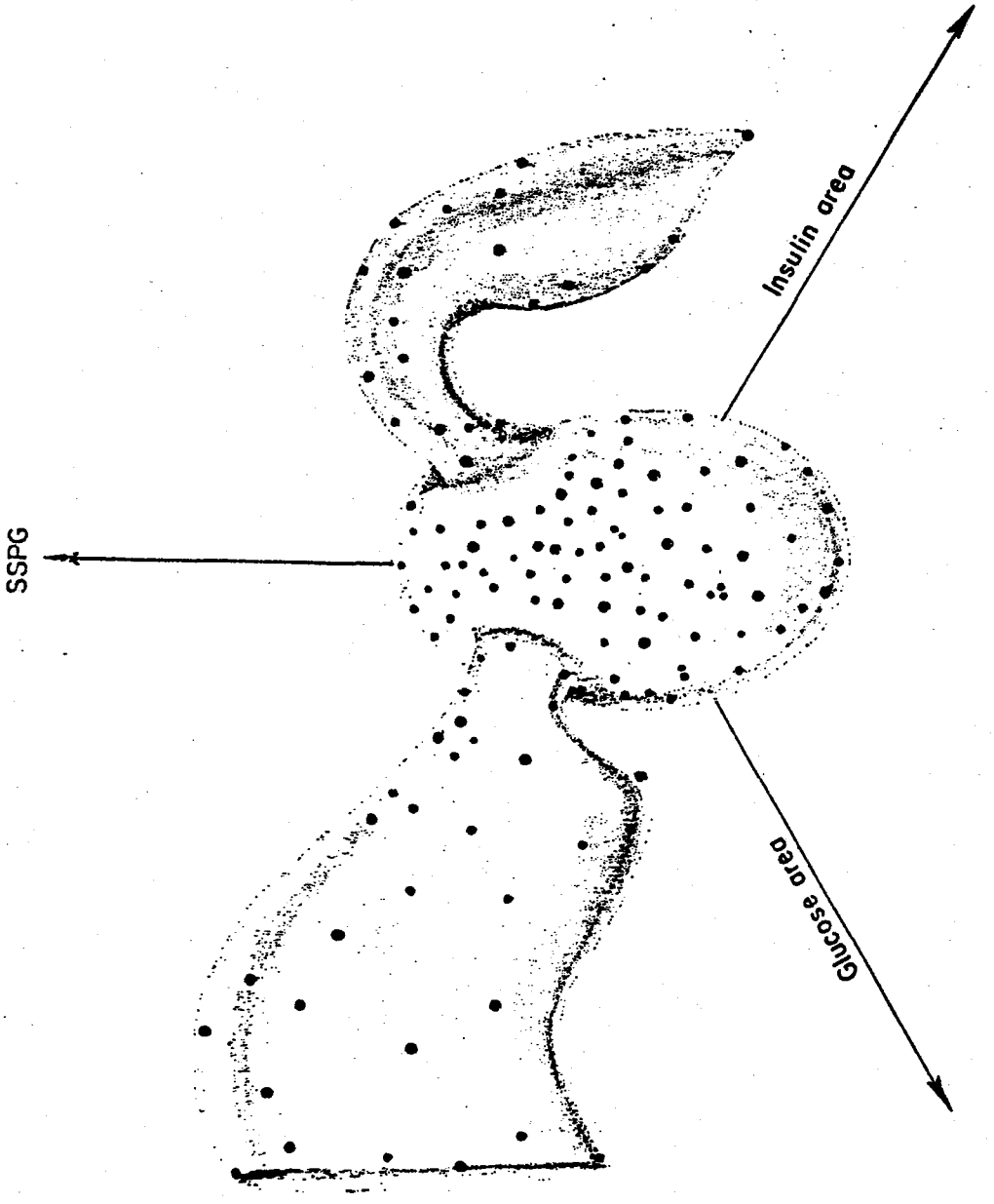

TABLE 1. CLUSTERING OF REAVEN AND MILLER'S DIABETES TRI-VARIATE DATA INTO THREE GROUPS

Classification Scheme	Group*			Criterion Value
	Normals (76,0,0)	Chemical Diabetes (0,36,0)	Overt Diabetes (0,0,33)	
Variant of $\|W\|$ with a priori cluster means specified Reaven and Miller (1977)	(73,10,1)	(3,26,6)	(0,0,26)	not given
Determinant of within Groups Sum of squares, $\|W\|$. See (7).	1: (76,0,0)	(0,36,0)	(0,0,33)	0.2217×10^{19}
	2: (73,17,3)	(3,19,4)	(0,0,26)	0.7416×10^{18}
	3: (73,17,3)	(3,19,4)	(0,0,26)	0.7416×10^{18}
	4: no other local minima found			
Maximum likelihood modification of $\|W\|$. See (8).	1: (76,0,0)	(0,36,0)	(0,0,33)	0.4978×10^{4}
	2: (75,30,6)	(1,6,1)	(0,0,26)	0.4750×10^{4}
	3: (75,30,6)	(1,6,1)	(0,0,26)	0.4750×10^{4}
	4: (76,36,7)	(0,0,11)	(0,0,15)	0.4754×10^{4}
Bayesian modification of $\|W\|$. See (10).	1: (76,0,0)	(0,36,0)	(0,0,33)	0.5182×10^{4}
	2: (75,30,6)	(1,6,1)	(0,0,26)	0.4951×10^{4}
	3: (76,30,6)	(1,6,1)	(0,0,26)	0.4951×10^{4}
	4: (76,36,7)	(0,0,11)	(0,0,15)	0.4953×10^{4}
Metric criterion with $\|W_g\|$. See (13).	1: (76,0,0)	(0,36,0)	(0,0,33)	0.8141×10^{6}
	2: (73,10,0)	(3,26,7)	(0,0,26)	0.6123×10^{6}
	3: (73,10,0)	(3,26,7)	(0,0,26)	0.6123×10^{6}
	4: (73,12,1)	(3,24,6)	(0,0,26)	0.6134×10^{6}
Maximum likelihood criterion with $\|W_g\|$. See (11).	1: (76,0,0)	(0,36,0)	(0,0,33)	0.4225×10^{4}
	2: (63,0,0)	(13,30,2)	(0,6,31)	0.4128×10^{4}
	3: (63,0,0)	(13,30,2)	(0,6,31)	0.4128×10^{4}
	4: (68,5,0)	(7,25,2)	(1,6,31)	0.4130×10^{4}
Bayesian criterion with $\|W_g\|$. See (12).	1: (76,0,0)	(0,36,0)	(0,0,33)	0.3095×10^{4}
	2: (73,9,0)	(3,27,5)	(0,0,28)	0.3008×10^{4}
	3: (72,9,0)	(4,23,2)	(0,4,31)	0.300457×10^{4}
	4: (73,9,0)	(3,27,7)	(0,0,26)	0.300528×10^{4}

*This division into three groups of the 145 patients was given in Reaven and Miller (1977) based upon Glucose Tolerance Test.

°The four table entries are 1: classification and criterion value for the clinical classification; 2: classification with minimum criterion value found by starting at the clinical classification; 3 and 4: classification with minimum criterion value for the two smallest criteria values produced by McRae's (1971) program.