J. Statist. Comput. Simul., 1980, Vol. 11, 87–106
0094-9655/80/1102-0087 $4.50/0

Printed in Great Britain

A Simulative Comparison of Linear, Quadratic and Kernel Discrimination

J. REMME, J. D. F. HABBEMA and J. HERMANS†

Department of Medical Statistics, University of Leiden, Wassenaarseweg 80, P.O. Box 9512, 2300 RA Leiden, The Netherlands

(*Received July 16, 1979*)

Three discriminant analysis models: linear discriminant analysis, quadratic discriminant analysis and a model using kernel estimation, are compared in a simulation study. Data are generated from multinormal, lognormal and mixtures of distributions. Some measures of performance that are based upon posterior probabilities are used for the comparison. The influence on performance of sample size, dimensionality and distance between populations is investigated.

An algorithm to estimate the smoothness parameters for the kernel model is evaluated. Finally some remarks are made with respect to the so-called variable kernel model.

KEY WORDS: Posterior probabilities, linear discriminant analysis, quadratic discriminant analysis, kernel model, variable kernel model, smoothness parameters.

1. INTRODUCTION

In this paper we will limit ourselves to Bayes' discriminant analysis. This approach is based upon posterior probabilities, as calculated by Bayes' theorem. These probabilities depend upon the probability density func-

---

†All authors were in the Department of Medical Statistics, University of Leiden, Wassenaarseweg 80, P.O. Box 9512, 2300 RA Leiden. J. Remme is now lecturer in the Department of Epidemiology and Biostatistics, University of Dar-es-Salaam, Tanzania; J. D. F. Habbema is now in the Department of Public Health and Social Medicine, Erasmus University of Rotterdam. The authors wish to express their appreciation to Mrs. E. Arents-Bos for the accurate typewriting of the manuscript and to Dr. M. E. Wise for help with the English text.

tions (pdf), which are usually estimated. Several statistical models may be considered for the density estimation. However, different models may lead to different estimated pdf's and hence to quite different posterior probabilities (Hermans and Habbema, 1975).

An attractive way to study this is by means of a simulation study. Simulation gives the opportunity to calculate the true posterior probabilities, using the known densities. The performance of a discriminant analysis model can then be evaluated by examining the agreement between the posterior probabilities as estimated under this model and the true ones.

In this paper we present the results of a simulation study designed to compare the performance of three important discriminant analysis models (see Section 2), which result from estimating the densities from

- multinormal distributions with equal covariance matrices (Linear Discriminant Analysis)
- multinormal distributions with unequal covariance matrices (Quadratic Discriminant Analysis)
- direct density estimation by means of kernel functions (Kernel model).

The simulation study is restricted to continuous data and to discrimination between two populations only. The data are generated from multinormal, lognormal and mixtures of multinormal distributions. Emphasis is on small sample sizes. Several measures for agreement between true and estimated posterior probabilities will be discussed in Section 3.2.

Evaluation of discriminant analysis models are given by Gilbert (1968), Marks and Dunn (1974), Goldstein (1975), Van Ness and Simpson (1976) and Aitchison *et al.* (1977). Only the paper by Van Ness and Simpson includes a comparison of models considered in the present paper. However, their study is limited to data generated from multinormal distributions with equal covariance matrices.

Kernel functions involve so-called smoothness parameters. In the study by Van Ness and Simpson the smoothness parameters are estimated using extra generated data. This surplus information is never available in applications. In the present study we use for the estimation of the smoothness parameters an algorithm which only needs the data on which the density estimation has to be based (see Section 2.3). This algorithm will be discussed critically in Section 5. Finally, the so-called variable kernel model, as mentioned by Victor (1976) and considered in more detail by Breiman *et al.* (1977), and Habbema *et al.* (1978), will be briefly discussed.

2. THE BAYES DISCRIMINANT ANALYSIS PROBLEM AND THE THREE MODELS

In discriminant analysis problems with two populations Π_1 and Π_2 ($\Pi_1 \cap \Pi_2 = \emptyset$) one has to assess for an element X ($X \in \Pi_1 \cup \Pi_2$) with a given p-dimensional observation x, the relative plausibilities for Π_1 and Π_2. The appropriate tools for the quantification of these plausibilities are the posterior probabilities $P(\Pi_i | x)$ ($i=1,2$) as calculated by Bayes' theorem:

$$P(\Pi_i | x) = P(\Pi_i) \cdot f(x | \Pi_i) \Big/ \sum_{j=1}^{2} P(\Pi_j) \cdot f(x | \Pi_j) \qquad i=1,2 \qquad (2.1)$$

where the $P(\Pi_i)$ are the prior probabilities.

When the objective is to allocate x to one of the populations, allocation rules are used which are based upon these posterior probabilities. The forced allocation rule is well known: allocate X to the population with the highest posterior probability.

In applications the priors $P(\Pi_i)$ and the pdf's $f(x | \Pi_i)$ are usually not known but have to be estimated. This estimation gives, using (2.1), estimated values for $P(\Pi_i | x)$.

Each pdf $f(x | \Pi_i)$ has to be estimated using the p-dimensional observations on sample elements from the corresponding population, the so-called training samples (of sizes N_1 and N_2). We shall indicate the jth training element from population Π_i by Y_{ij} and the corresponding observation by y_{ij}.

In the three models under consideration the density estimation is based upon:

- Multinormal distributions with equal covariance matrices
  This gives rise to a linear discriminant analysis (LDA) which is frequently applied, e.g. in BMDP (see Dixon, 1975) and SPSS (see Nie, 1975). The pdf's $f(x | \Pi_1)$ and $f(x | \Pi_2)$ are estimated from the equation

$$\hat{f}_{\mathrm{LDA}}(x | \Pi_i) = N^{(p)}(x | \bar{y}_i, S) \qquad (i=1,2) \qquad (2.2)$$

  where $\bar{y}_i$ is the mean of the training sample from Π_i and S is the pooled sample covariance matrix.
- Multinormal distributions with unequal covariance matrices
  This leads to a quadratic discriminant analysis (QDA). The pdf's are estimated from

$$\hat{f}_{\mathrm{QDA}}(x | \Pi_i) = N^{(p)}(x | \bar{y}_i, S_i) \qquad (i=1,2) \qquad (2.3)$$

where $\bar{y}_i$ again is the mean and S_i the covariance matrix for the training sample from Π_i.

– Kernel functions (K)

A kernel function $K_i(x \mid y_{ij})$ with the properties of a pdf is laid around each point y_{ij} of the training sample from population Π_i. The pdf $f(x \mid \Pi_i)$ is then estimated from

$$\hat{f}_K(x \mid \Pi_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} K_i(x \mid y_{ij}) \tag{2.4}$$

In our simulation study we chose a multinormal kernel with uncorrelated coordinates, which is also used in the ALLOC programs for discriminant analysis, see Hermans and Habbema (1976).

$$K_i(x \mid y_{ij}) = \frac{1}{(2\pi)^{p/2} \mid H_i \mid^{1/2}} \exp\{-\tfrac{1}{2}(x-y_{ij})^T H_i^{-1}(x-y_{ij})\} \tag{2.5}$$

where

$$H_i = h_i^2 \begin{pmatrix} s_{i1}^2 & 0 & \dots & 0 \\ 0 & s_{i2}^2 & & \\ \vdots & & & \vdots \\ 0 & \dots & & s_{ip}^2 \end{pmatrix} \qquad i = 1, 2$$

and s_{it}^2 $(t = 1, p)$ are the variances in the training sample from Π_i. More details about the use of kernel functions in discriminant analysis may be found in Meisel (1972).

A crucial point in the use of kernel models is the estimation of the so-called smoothness parameter h_i. Trying to use the information from the training data, one might think of a pseudo maximum likelihood method: find the value $\hat{h}_i$ which maximizes

$$L_i(h) = \prod_{j=1}^{N_i} \hat{f}_K(y_{ij} \mid \Pi_i) \tag{2.6}$$

However, this leads to the degenerate result $\hat{h}_i = 0$. This degeneration can be avoided by using a jack-knife or leaving-one-out modification of (2.6), as proposed by Habbema *et al.* (1974a) and by Duin (1976): $\hat{h}_i$ is the maximizing value of $LL_i(h)$, with

$$LL_i(h) = \prod_{j=1}^{N_i} \hat{f}_K^j(y_{ij} \mid \Pi_i) \tag{2.7}$$

where $\hat{f}_K^i(x \mid \Pi_i)$ is the kernel density estimate based on all the sample elements from Π_i except Y_{ij}. In our study we have used this LOO method for the estimation of h_i.

Using the density estimates defined in (2.2), (2.3) and (2.4) the posterior probabilities estimated under each model M ($M = \text{LDA}, \text{QDA}, \text{K}$) are calculated from

$$P_M(\Pi_i \mid x) = P(\Pi_i)\hat{f}_M(x \mid \Pi_i) \Big/ \sum_{j=1}^{2} P(\Pi_j)\hat{f}_M(x \mid \Pi_j) \qquad i = 1, 2 \qquad (2.8)$$

3. DESIGN OF THE SIMULATION EXPERIMENT

Several combinations of the input parameters, namely the underlying distribution of the data, sample sizes and dimensionality, have been chosen in the simulation study. In each simulation, training data were generated to estimate the densities according to each of the three models. Furthermore, additional data, the so-called test data, were generated. For these test data the estimated posterior probabilities $P_M(\Pi_i \mid x)$ and the true posterior probabilities $P(\Pi_i \mid x)$ were calculated. The priors were always taken to be equal, and all simulations were run with only two populations.

3.1 Chosen input for the simulations

Characteristic for many applied problems is the small size N_i ($i = 1, 2$) and a not necessarily small dimension p. For this reason most simulations were run with 15 or 35 training elements per population ($N_1 = N_2 = 15$ or 35), while the number of variables p varied from 2 to 10. The data were generated from members of the following classes of distributions:

EQ—multinormal distributions with equal covariance matrices
UNEQ—multinormal distributions with unequal covariance matrices
LOG—multidimensional lognormal distributions
MIX—mixtures of multinormal distributions.

The data were obtained by transforming standard normally distributed random numbers, generated with a subroutine from the NAG package (see NAG, 1976).

We shall use the term "simulation" for a particular combination of input parameters. For each simulation 25 runs were executed. In such a run N training and 50 test observations were generated from each

for appraising the estimates of posterior probabilities than error rates; the remaining measures are based upon this idea. Since we simulated with only two populations the comparison can be simplified by just looking at the posterior probability for the first population Π_1.

The first one of these measures is the percentage of test elements allocated to the same population with both $P(\Pi_1 | x)$ and $P_M(\Pi_1 | x)$ using the forced allocation rule. The next two measures are based on the allocation that includes "doubt". With again a doubt threshold of $\delta = 0.90$ each test element was allocated both with $P(\Pi_1 | x)$ and $P_M(\Pi_1 | x)$; according to the combination of these allocations the test element was classified into one of the 9 classes in Table I.

TABLE I

Cross-tabulation of true posterior probabilities, $P(\Pi_1 | x)$, and estimated ones, $P_M(\Pi_1 | x)$. The boxes a_i correspond to "agreement", e_i to "slight error" and d_i to "serious disagreement"

$P_M(\Pi_1 \mid x)$				
[0.9–1.0]	d_1	e_1	a_3
(0.1–0.9)	e_4	a_2	e_2
[0.0–0.1]	a_1	e_3	d_2
	[0.0–0.1]	(0.1–0.9)	[0.9–1.0]
		$P(\Pi_1 \mid x)$	

The two measures derived from Table I are:

measure A: the percentage of test elements which fall into classes a_1, a_2 or a_3.

measure D: the percentage of test elements which fall into classes d_1 or d_2. (3.2)

The measure A is an index for the agreement while D measures serious disagreement between the true posterior probabilities and the estimated ones, using model M.

For the presentation in this paper we restrict ourselves to the measures A and D since these measures together give a reasonably good impression of the overall picture. For more refined measures, based on the exact values of P and P_M, see Hilden *et al.* (1978).

3.3 Measure for the separability of the populations

Separability is in most simulation studies characterized by the Mahalanobis distance, which is based on the assumption of multinormality with equal covariance matrices. But we generated the data not just from multinormal but also from skew and multimodal distributions, where the Mahalanobis distance is not appropriate. For this reason we used

measure C: the percentage of test elements for which the true posterior probability for the population of origin exceeds 0.90. (3.3)

The actual values for C are realizations of a random variable but these values are in each simulation based upon 2500 test samples; their standard errors are always smaller than 1 %.

4. RESULTS

A typical example of the results obtained is given in Figure 1. The data for Π_1 are generated from $N^{(2)}(0, I)$ and for Π_2 from $N^{(2)}(\mu_2, 4I)$ with $\mu_2^T = (\mu, 0)$. On the horizontal axis the values are plotted for the separability measure C, see (3.3). An increasing value for C corresponds to an increasing value for μ, see the second horizontal axis. For 7 values of μ data are generated and the LDA, QDA and K results are calculated.

The values for the performance measure A are plotted on the vertical axis. These values are means over 25 runs. To indicate the variability, the standard errors of the mean (s.e.m.) are given in Figure 1.

For all subsequent figures only the mean values are given in order to avoid a confusing amount of information in each figure. The maximal s.e.m.'s for the A scores in these figures were obtained with the lognormal distributions; for the LDA, QDA and Kernel model they were 3.4, 4.4 and 4.7 respectively. The respective maximal values in the other simulations were 2.1, 2.4 and 2.8. In most cases, they were much lower.

Results for the D measure are omitted when all values obtained for all the three methods are below 2 %. For the simulations reported in Figure 1 this was the case.

4.1 Multinormal distributions (EQ and UNEQ)

The data for Π_1 are generated from $N^{(p)}(0, I)$ and the data for Π_2 from $N^{(p)}(\mu_2, \Sigma_2)$ with $\mu_2^T = (\mu, 0, \ldots, 0)$ and $\Sigma_2 = \lambda \cdot I$. We examined problems in 2 and 6 dimensions ($p = 2, 6$); in Figure 2 only the results with $p = 6$ are given.

For equal covariance matrices ($\lambda=1$; Figure 2a and b) the best results were obtained with the LDA, as might be expected. These results are uniformly better than those with the QDA, while the K results are good for small differences (small C values) but become the worst of the three models for higher C values. For unequal covariance matrices ($\lambda=4$ and $\lambda=16$) the LDA had the worst performance; also serious errors (D-measure) are now present, see Figure 2e, f, i and j. Although in these cases the distributional assumptions for the QDA model hold, this model was not uniformly the best one. For $N=15$ the K model did better, while for $N=35$ the two are nearly equivalent. Note that the kernel model yielded no gross errors.

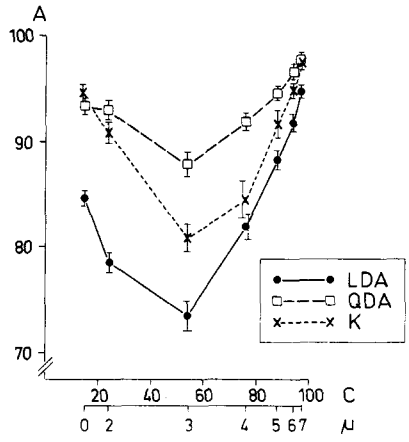

FIGURE 1. An example of results for multinormal distributions.

Note: The data are generated from $N^{(2)}(0, I)$ and $N^{(2)}(\mu_2, 4I)$ with $\mu_2^T=(\mu, 0)$. At each input value for μ a training sample of size 35 and a test sample of size 50 is generated from each population. To each μ value corresponds a C value based on 2500 test elements. The mean and standard error of the mean for the A measure are calculated for each of the three discriminant analysis methods. They are based on 25 runs with 100 test elements in each run.

Briefly with $p=2$ LDA was very bad for $\lambda=16$; for other values of λ there were no substantial differences between the three models; however, LDA was somewhat better for $\lambda=1$, and QDA for $\lambda=4$.

The data in the simulations with multinormal distributions were generated from uncorrelated distributions. This might favour the Kernel model (2.5) since uncorrelated multinormal kernel functions are used. In order to investigate to what extent correlations would change our conclusions, we repeated several of the multinormal simulations after linear transformations, that resulted in true multinormal distributions with correlations. It is well known that these transformations do not affect the

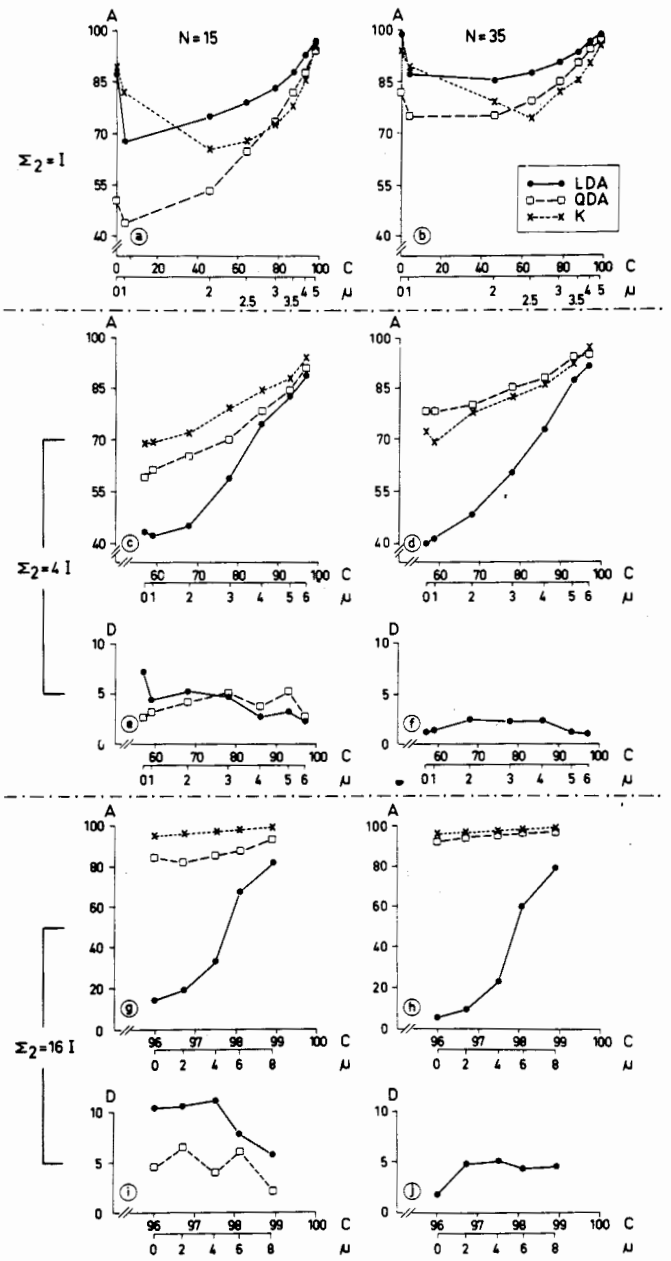

FIGURE 2. Results for six dimensional multinormal distributions.

performance of the LDA and the QDA (see e.g. Marks and Dunn, 1974). The transformation to multinormal distributions with correlations from 0.4 to 0.5 resulted for the Kernel model in a 1 to 5 units decrease of measure A (see also Remme *et al.*, 1977). The values for measure D did not change appreciably. These small differences hardly affect the overall picture in this study. Some of the simulations with $p=6$ and all correlations over 0.6 showed a more serious decrease of the performance of the Kernel model. Therefore the results of this study should be handled with caution when there is too much correlation between the variables.

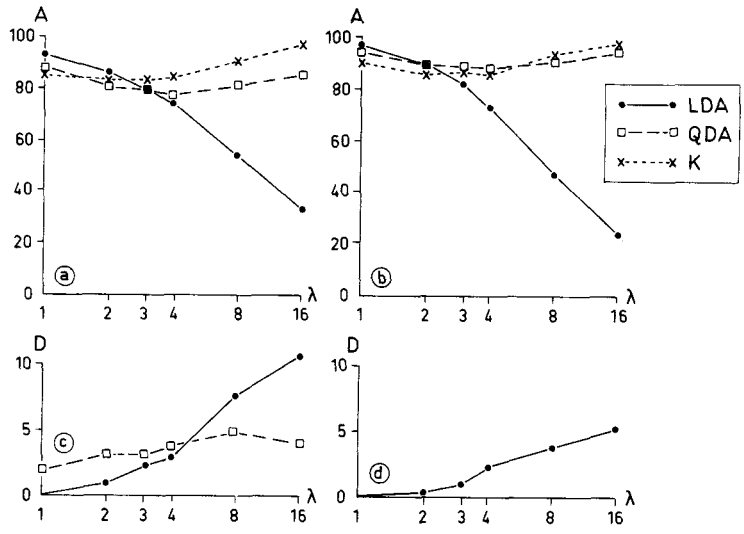

FIGURE 3. Influence of unequal covariance matrices.

Note: Data are generated from $N^{(6)}(0, I)$ and $N^{(6)}(\mu_2, \lambda I)$ with $\mu_2^T = (\mu, 0, \ldots, 0)$. The μ values were chosen such that the corresponding C values were about 90%.

The influence of unequal variances is studied in more detail. Figure 3 gives the results for $p=6$. The scores on measure A for the QDA are about equal to the LDA scores for $\lambda=2$. For $\lambda=3$ the performance of the LDA model is in all figures the worst and this performance deteriorates rapidly for increasing λ. These results indicate a serious restriction for the use of the LDA. The factor λ hardly influences the difference in performance between the QDA and the Kernel model. For $p=2$ the results for the A measure were very similar to $p=6$; all D values were substantially lower.

4.2 Lognormal distributions (LOG)

The data in simulations with lognormal distributions were obtained from multinormal distributed random numbers by the following transformations:

$$X_1 \in \Pi_1 : x_1 = \exp(v_1) \quad \text{with} \quad v_1 \sim N^{(2)}(\mu_1, \Sigma_1)$$
$$X_2 \in \Pi_2 : x_2 = b + \exp(v_2) \quad \text{with} \quad v_2 \sim N^{(2)}(\mu_2, \Sigma_2) \tag{4.1}$$

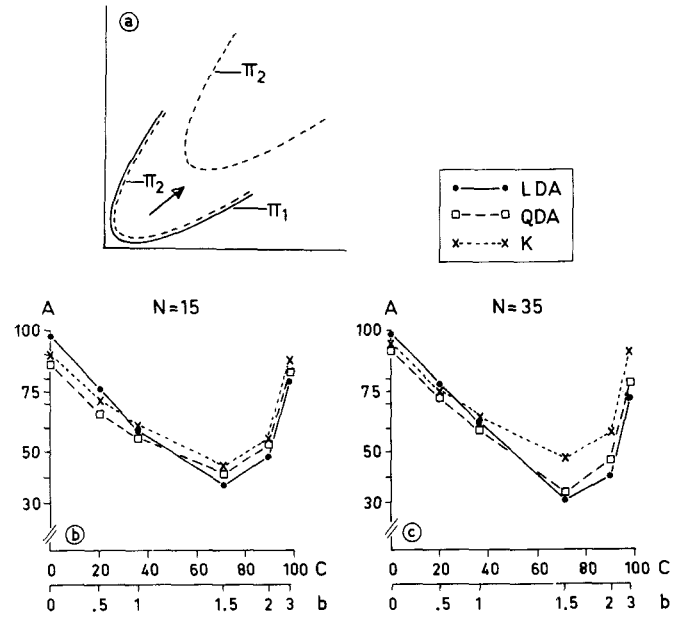

FIGURE 4. Results for lognormal distributions.

Note: The data are generated according Eq. (4.1) with $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$, $\mu_1 = \mu_2 = (0, 0)$ and the translation vector (b, b) for Π_2 ranges from $b = 0$ to $b = 3$.

Figures 4a and 5a show density contours of the true distributions used. Increasing values for measure C correspond to increasing values for the translation vector b, ranging up from $(0, 0)$ to $(3, 3)$ for 4a and to $(5, 0)$ for 5a. The density contours for Π_2 are only plotted for the two extreme situations. A comparison of the results in Figures 4 and 5 with those in Figure 2 shows that in all models it is more difficult to estimate the posterior probabilities for these lognormal distributions than for multinormal distributions. The Kernel model is only a little better in Figure 4c and the QDA model profits slightly from the different covariance structure in

Figure 5b and c. Seriors errors, up to just 2%, were equally present in all three models.

The poor results for the LDA and the QDA were to be expected because of the violation of their distributional assumptions. The kernel model is distribution free and therefore better results might be expected from it. However, we will come back to this aspect in Section 5.2.

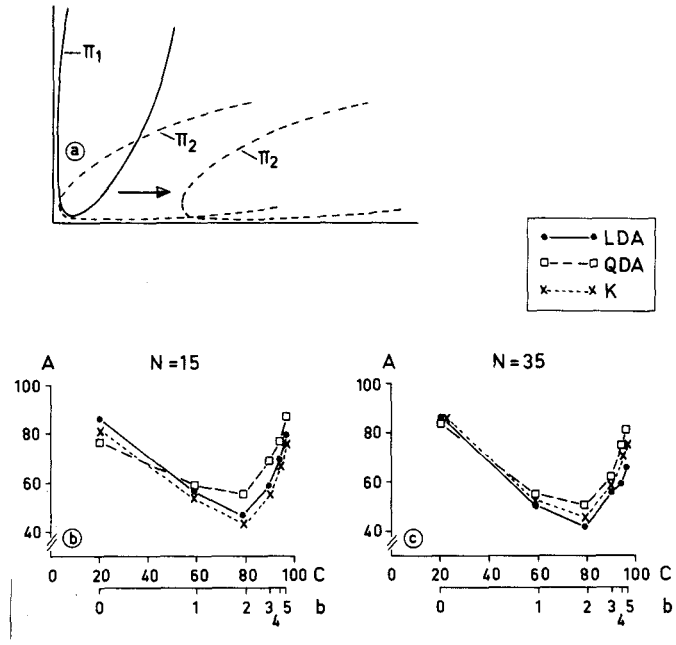

FIGURE 5. Results for lognormal distributions.

Note: The data are generated according Eq. (4.1) with $\Sigma_1 = \begin{pmatrix} 0.65 & 0.5 \\ 0.5 & 0.9 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.9 & 0.5 \\ 0.5 & 0.65 \end{pmatrix}$, $\mu_1 = (0.4, 0.9)$, $\mu_2 = (0.9, 0.4)$ and the translation vector $(b, 0)$ for Π_2 ranges from $b=0$ to $b=5$.

4.3 Mixtures of multinormal distributions (MIX)

Each mixture component is a multinormal distribution with the identity matrix I as covariance matrix. The mixing proportions for the components are taken to be equal. Simulations were run for different values of b, the translation vector for Π_2. Figure 6a shows for one of the examined situations density contours of the true distributions for Π_1 and Π_2. For Π_2 again the contours for the two extremes are shown.

With respect to the A measure the LDA and QDA models are equivalent, while the Kernel model is superior. There were very few

serious errors with any model. Simulations with Π_1 bimodal and Π_2 unimodal gave no substantial differences between the three models.

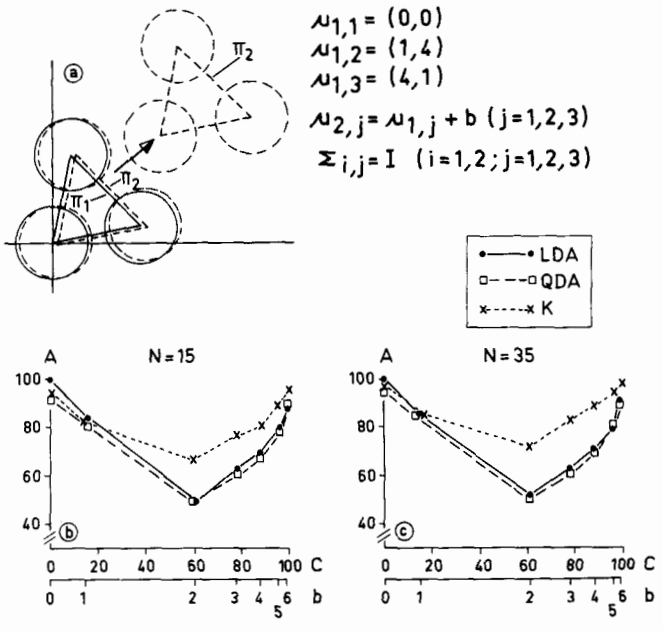

FIGURE 6. Results for mixtures.

Note: The translation vector (b, b) for Π_2 ranges from $b=0$ to $b=6$.

4.4 Influence of the number of training samples (N)

Our main objective is to study the performance of the models for small sample sizes. In this section, however, we shall give results for increasing sample sizes up to 100 or 200 samples per population.

Figure 7 gives some results. For six-dimensional multinormal distributions with unequal covariance matrices the LDA shows no improvement for the A measure with increasing N, the QDA starts with a bad performance but gives a sharp improvement, while the Kernel model starts well and gives a steady improvement. With respect to the D measure the Kernel model is remarkably good. For the lognormal and mixture distributions the Kernel model soon becomes superior to the LDA and QDA. The latter two show of course for these distributions no improvement with increasing N. For the lognormal distributions the behaviour of the Kernel model remains unsatisfactory. In the next section we shall take up this point again.

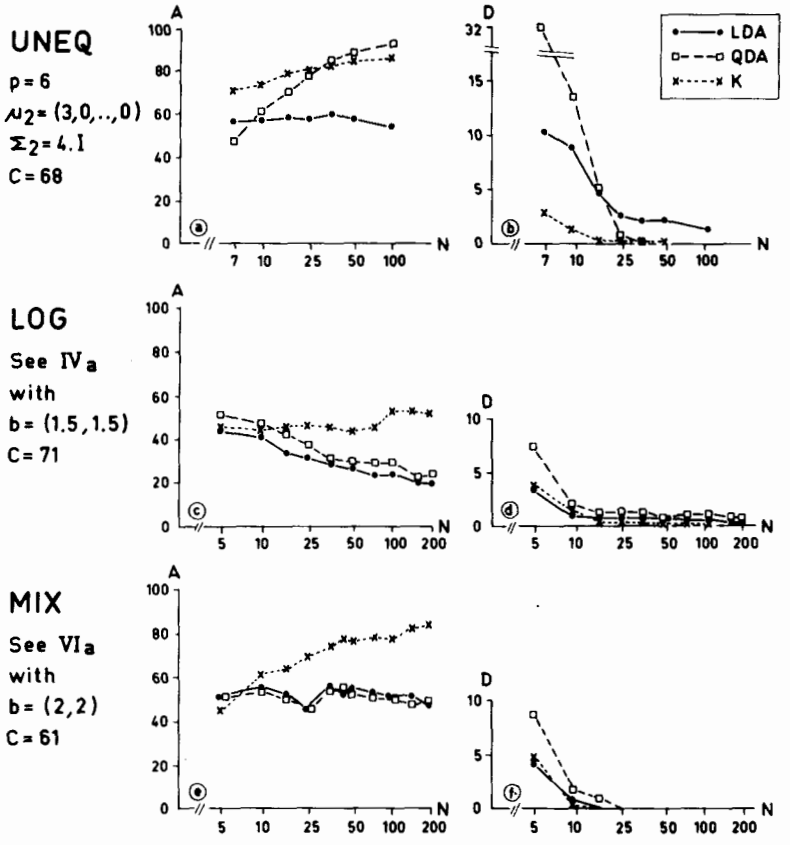

FIGURE 7. Influence of the size N of the training samples.

5. THE ESTIMATION OF SMOOTHNESS PARAMETERS IN THE KERNEL MODELS

In the use of kernel density estimates the choice of the smoothness parameters h_i, see (2.5), is crucial. One of the first attempts to find a good estimation procedure for these parameters was made by Rosenblatt (1956) for the univariate case. He suggested taking $h_i = \alpha_i N_i^{-1/5}$, with α_i dependent on the true pdf and its second derivative. This true density is never known in applications, and it seems to be necessary to let the sample data play some role in estimating h_i. This approach has been used in the LOO estimation method (2.7). However, it is still uncertain whether these LOO estimates for h are appropriate for estimating the posterior probabilities.

5.1 Evaluation of the leaving-one-out (LOO) estimation method

To investigate the performance of this LOO method we ran some simulations (again with 25 runs) with input values for $h=h_1=h_2$ instead of having h_1 and h_2 calculated by the LOO method. The results are shown in Figure 8 with, on the horizontal axis, the chosen input values for h and, on the vertical axis, the performance measure A. The same simulations were run once more, but now with h_1 and h_2 estimated by the

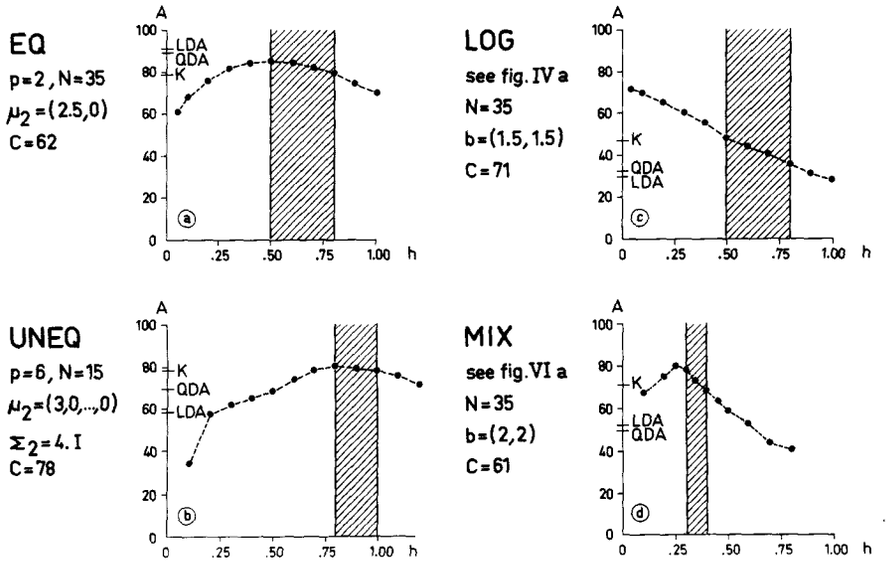

FIGURE 8. Effect of the smoothness parameter h.

Note: The curve ●—● indicates the A values, averaged over 25 runs, obtained with input values $h=h_1=h_2$ for the smoothness parameter in both populations. Exactly the same simulations were run with h_1 and h_2 estimated by the LOO method. The shaded area indicates on the horizontal axis the range of estimated h values in these 25 runs. The average A values for LDA, QDA and K are marked on the vertical axis.

LOO method. The shaded areas in the figures indicate the range of these estimated h-values in the 25 runs. The A values, averaged over the 25 runs for the LDA, the QDA and the Kernel, are indicated on the vertical axis.

For the multinormal distributions (EQ and UNEQ) and the mixture (MIX) the range of h-values resulting from the LOO method is satisfactory, although a little high. These high h-values result, via smoothed kernels, in rather smoothed densities, thus leading to prudent estimates for the posterior probabilities.

In the lognormal case (Figure 8c) the LOO algorithm seems to give bad results: substantially higher A values are obtained for h-values which are markedly lower than those obtained with the LOO method. However, taking also into account the serious errors (D-measure) it turns out that these errors also increase sharply with decreasing h-values. So for more definite conclusions, an overall performance measure of the type $\alpha A - \delta D$ should be considered.

In contrast to the other three figures there is for the lognormal situation no clear optimum for h and one wonders whether the present kernel model is appropriate for lognormal distributions!

5.2 Variable kernel model

So far we have used only one smoothness parameter h_i within each population. All kernels in a density estimate will then have the same shape. This might be a disadvantage for skew distributions, because kernels around isolated points in the tail will be too peaked and kernels in high density regions too smoothed.

A model which avoids these problems is the so-called variable kernel model (Breiman *et al.*, 1977). They do not estimate just one smoothness parameter h_i for each density estimate $f_K(x \mid \Pi_i)$. For each kernel $K(x \mid y_{ij})$ a smoothness parameter h_{ij} is taken, which depends not only on the population Π_i, but also on the observation point y_{ij}:

$$h_{ij} = \alpha_i d_{ijk_i} \tag{5.1}$$

where d_{ijk_i} is the distance from y_{ij} to its k_i-th nearest neighbour in Π_i and α_i is the overall smoothness parameter for the density estimate for Π_i. Now two parameters α_i and k_i have to be estimated for each population. Breiman *et al.* (1977) studied two bivariate examples. In these two examples they found a relation between α_i and k_i which left them only one parameter to estimate. However, in our simulations with $N = 15$ and $N = 35$ we could not reproduce this relation, and therefore we used another approach in the simulations in which we evaluated the variable kernel model. The results are published in Habbema *et al.* (1978). Recalling its conclusions it can be stated that in the LOG case the variable kernel is superior to those for the (non-variable) kernel model. For EQ, UNEQ and MIX there are no substantial differences between the results for the kernel model and the variable kernel model. However, the performance of the variable kernel model depends on the choice of k_1 and k_2, and a good algorithm for the estimation of the k_i's is not yet available. This prohibits its routine use, in spite of its promising results.

6. DISCUSSION

As might be expected the LDA model showed better results than the two other models in the simulations with multinormal distributions with equal covariance matrices. However, for unequal covariance matrices ($\Sigma_1 = I$ and $\Sigma_2 = \lambda I$) the performance of the LDA model deteriorated for increasing λ and was very poor for $\lambda > 4$. The QDA model was only slightly better than the other models in some simulations with two-dimensional multi-normal distributions with unequal covariance matrices. Decisive for the performance of this model were the sample sizes in relation to the dimensionality. The QDA is quite inappropriate for problems with relatively small sample sizes, as already shown in the excellent paper of Marks and Dunn (1974). The measures as used in our study showed this failure even more clearly. Aitchison *et al.* (1977) pointed out that the performance of the LDA and the QDA in very small sample problems might be improved by using the Bayesian or predictive approach. However, we have the impression from a few extra runs that this improvement is rather small for our simulations and that it would not change the main trends in our results.

The performance of the Kernel model was better than or about equal to the performance of the other models, except in the simulations with multinormal distributions with equal covariance matrices. It was surprising that the Kernel model in relation to the other model showed very good results in problems with relatively small samples. Possibly this was because of the use of the leaving-one-out estimation method for the smoothness parameters, which seems to be an advantage in small sample problems. This method turned out to give good but rather prudent posterior probability estimates. The kernel results for the lognormal case were quite disappointing. A much better performance was here obtained for the variable kernel model, see Section 5.2. The results for this variable kernel model in simulations with multinormal distributions and with mixtures were about equal to those for our original kernel model. However, the variable kernel model involves an extra parameter and a good algorithm for its estimation is not yet available.

Some simulations were also run with sample sizes increasing up to 100 or 200 per population. The Kernel model performed increasingly well with increasing sample sizes in all these simulations because of its distribution-free nature. However, this improvement was quite slow in the lognormal case. For the LDA and the QDA increasing the sample sizes yielded better results only when the data were generated from multinormal distributions.

It must be born in mind that the results and conclusions in this study are limited to discriminant analysis problems for which not all variables are highly correlated, say to problems in which not all correlations are higher than 0.5. A second restriction of this study is that only simulations with two populations were run. The effect of both high correlations and more than two populations is under investigation.

Finally, we would like to recommend the following on the use of the three models. For applications the Kernel model is available. Although each of the models yields the best results in some situations, the Kernel model is the only one which was the best or close to the best in all situations we considered. In particular the present practice of the nearly exclusive use of the LDA cannot be justified by our results.

References

Aitchison, J., Habbema, J. D. F. and Kay, J. (1977). A critical comparison of two methods of statistical discrimination. *Applied Statistics* **26,** 15–25.

Breiman, L., Meisel, W. and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics* **19,** 135–144.

Dixon, W. J. (ed.). (1975). *BDMP, Biomedical Computer Programs,* University of California Press. Berkeley and Los Angeles.

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers* **C-25,** 1175–1179.

Gilbert, E. S. (1969). The effect of unequal variance–covariance matrices on Fisher's linear discriminant function. *Biometrics* **25,** 505–515.

Goldstein, M. (1975). Comparison of some density estimate classification procedures. *JASA* **70,** 666–669.

Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974a). A stepwise discriminant analysis program using density estimation. *COMPSTAT 1974, Proceedings in Computational Statistics.* Physica Verlag, Wien.

Habbema, J. D. F., Hermans, J. and van der Burgt, A. T. (1974b). Cases of doubt in allocation problems. *Biometrika* **61,** 313–324.

Habbema, J. D. F., Hermans, J. and Remme, J. (1977). Data analytical methods in discriminant analysis: The analysis of posterior probabilities. In: *Data Analysis and Informatics.* IRIA Symposium. Rocquencourt. 211–221.

Habbema, J. D. F., Hermans, J. and Remme, J. (1978). Variable kernel density estimation in discriminant analysis. *COMPSTAT 1978, Proceedings in Computational Statistics.* Physica Verlag, Wien.

Hermans, J. and Habbema, J. D. F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Medizin und Biologie* **6,** 14–19.

Hermans, J. and Habbema, J. D. F. (1976). The Alloc package for multigroup discriminant analysis programs based on direct density estimation. *COMPSTAT 1976, Proceedings in Computational Statistics.* Physica Verlag, Wien.

Hilden, J., Habbema, J. D. F., and Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis, III. Method based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine* **17,** 238–246.

Marks, S. and Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. *JASA* **69,** 555–559.

Meisel, W. S. (1972). *Computer-oriented Approaches to Pattern Recognition.* Academic Press, New York.

NAG (1976). *Manual for the NAG Package, Mark V,* Numerical Algorithms Group, 7 Banbury Road, Oxford.

Van Ness, J. W. and Simpson, C. (1976). On the effect of dimension in discriminant analysis. *Technometrics* **18,** 175–187.

Nie, N. H., Hadlai Hull, C., Jenkins, J. G., Steinbrenner, K. and Bent, D. H. (1975). *SPSS,* second edition. McGraw-Hill Book Comp., New York.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33,** 1065–1076.

Remme, J., Habbema, J. D. F. and Hermans, J. (1977). A comparison of the performance of two discriminant analysis models. In: *Data Analysis and Informatics,* IRIA Symposium, Rocquencourt, 269–279.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27,** 832–837.

Victor, N. (1976). Non-parametric allocation rules. In: *Decision Making and Medical Care* (eds.) F. T. de Dombal and F. Grémy. North-Holland Publishing Cie., Amsterdam.