ISBN 82-553-0736-2 No. 5
April 1991

Statistical classification using a linear mixture of two multinormal probability densities

by

Torfinn Taxt[a], Nils Lid Hjort[a,b]
and Line Eikvil[a]

Statistical classification using a linear mixture of two multinormal probability densities

Torfinn Taxt[a], Nils Lid Hjort[a,b] and Line Eikvil[a]

Norwegian Computing Centre[a] and University of Oslo[b]

[April 1991; to appear in 'Pattern Recognition Letters']

ABSTRACT. This paper describes an estimation-maximisation algorithm to estimate the parameters of a probability density model consisting of a linear mixture of two multinormal distributions. Superior classification results to those obtained using the multinormal distribution or the k-nearest-neighbour rule were obtained with this model on two difficult data sets.

KEY WORDS: *bimodality, classification, EM-algorithm, mixture distribution, multinormal, updating*

1. Introduction

Mixture distributions are an established statistical field (see Titterington, Smith and Makov (1985) for a recent, extensive review), but have so far been applied only to a limited extent in supervised statistical classification (but see Hosmer (1978) and Lachenbruch and Brofitt (1980) for the univariate case). Our interest in mixture distributions was initiated by the bimodal or multimodal empirical distributions that we frequently observed for feature vectors from the same class, when working with handwritten symbols and remote sensing images. This phenomenon was often observed both on the feature component level and in the multi-dimensional feature space, for several feature measurement methods.

With no specific limitations concerning the size of the training set and the classification rate, these types of multidimensional distributions could have been modelled and estimated nonparametrically. This typically places too high demands on the size of the training sets, however. In our industrial applications it was essential that the training set could be kept as small as possible, while maintaining a high correct classification rate. Taking these demands into account, we decided to model for each class these irregular and multidimensional feature vector distributions by a linear combination of two multinormal distributions. The estimation of the parameters in such distributions is known to be a difficult problem, both from the statistical and the algorithmic points of view; see for example Titterington, Smith and Makov (1985).

By applying a clustering technique to get good initial estimates for the Estimation-Maximisation (EM) algorithm we applied, we obtained a good fit of two different mixture models to a large majority of the unimodal and bimodal feature vector components we studied. Furthermore, estimation of these mixture functions demanded training sets of about the same size as the multinormal function to give the same or better results in statistical classification.

Parts of this study have been presented at a conference (Taxt, Eikvil and Hjort, 1989).

2. EM-algorithm for a linear mixture of two multinormal functions

To obtain numerical estimates of the parameters of the linear mixture models we followed Bezdek, Hathaway and Huggins (1985) and Hjort (1986) and derived iterative equations for the linear mixture of two multinormal functions using the EM approach to maximise the log likelihood (see Titterington, Smith and Makov, 1985, for example). We use $N_d(\mu, \Sigma)(x)$ to denote the familiar multinormal density, as a function of a d-dimensional x, with mean vector μ and covariance matrix Σ. *Model A* is defined by

$$f(x) = (1-\pi)N_d(\mu_1, K)(x) + \pi N_d(\mu_2, M)(x), \tag{1}$$

where x is the feature vector, μ_1 and μ_2 are the mean vectors, K and M the covariance matrices, and π the weight, $0 \leq \pi \leq 1$. *Model B* is defined by

$$f(x) = (1-\pi)N_d(\mu_1, S)(x) + \pi N_d(\mu_2, S)(x), \tag{2}$$

in which S is the covariance matrix common for both multinormal distributions. The other parameters are as defined for model A. We note that model A specifies no ties between the covariances of the two multinormal distributions which constitute the probability density model. In contrast, model B is more restrictive, and assumes that the two multinormal distributions have equal covariance matrices.

Of course there are many potentially interesting and useful models lying between model A and model B, utilising different structures of the covariance matrices, and one could similarly think of wider models with separate weights for separate sets of components. The present study is limited to models A and B, however.

Let the observed feature vectors of class c be $x_1, \ldots, x_n$ and define

$$Q^{(t)}(1|x) = (1-\pi^{(t)})f_1^{(t)}(x)/f^{(t)}(x) \quad \text{and} \quad Q^{(t)}(2|x) = \pi^{(t)}f_2^{(t)}(x)/f^{(t)}(x),$$

where $f_1(x)$ is either $N(\mu_1, K)(x)$ or $N(\mu_1, S)(x)$, $f_2(x)$ is either $N(\mu_2, M)(x)$ or $N(\mu_2, S)(x)$, $Q(1|x) + Q(2|x) = 1$, and t is indexing the number of iterations. The log-likelihood for model A given by

$$\log L = \log L(\pi, \mu_1, \mu_2, K, M) = \sum_{i=1}^{n} \log\{(1-\pi)N_d(\mu_1, K)(x_i) + \pi N_d(\mu_2, M)(x_i)\},$$

and there is a similar equation for model B. To describe the interative procedure that leads to numerical values for the maximising parameters, let first

$$n_0^{(t)}(1) = \sum_{i=1}^{n} Q^{(t)}(1|x_i) \quad \text{and} \quad n_0^{(t)}(2) = \sum_{i=1}^{n} Q^{(t)}(2|x_i).$$

In particular $n_0^{(t)}(1) + n_0^{(t)}(2) = n$. The following iterative EM equations emerge by setting partial derivatives of the log-likelihood equal to zero:

$$\pi^{(t+1)} = n_0^{(t)}(2)/n,$$

$$\mu_1^{(t+1)} = \frac{1}{n_0^{(t)}(1)} \sum_{i=1}^{n} Q^{(t)}(1|x_i)x_i, \quad \mu_2^{(t+1)} = \frac{1}{n_0^{(t)}(2)} \sum_{i=1}^{n} Q^{(t)}(2|x_i)x_i,$$

$$K^{(t+1)} = \frac{1}{n_0^{(t)}(1)} \sum_{i=1}^{n} Q^{(t)}(1|x_i)(x_i - \mu_1^{(t+1)})(x_i - \mu_1^{(t+1)})'$$

$$M^{(t+1)} = \frac{1}{n_0^{(t)}(2)} \sum_{i=1}^{n} Q^{(t)}(2|x_i)(x_i - \mu_2^{(t+1)})(x_i - \mu_2^{(t+1)})'.$$

Analogous iterative equations are found for model B. Some further details are in Hjort (1986, Ch. 3). It remains to define good starting values for these iterative equations. We obtained these starting values by clustering the n feature vectors of class c into two clusters using the 'k-means' partitional clustering algorithm (Anderberg, 1973, p. 162).

We have found it useful to incorporate a minor modification of this algorithm, consisting in running a simple component-wise test for one vs. two clusters first (Hjort, 1991). If the test signals one homogeneous class for component i, then the corresponding $\mu_{1,i}$ and $\mu_{2,i}$ components are set equal. This eliminates unnecessary parameters and helps convergence.

The full procedure is carried out for each class, giving estimates for π_c, $\mu_{1,c}$, $\mu_{2,c}$, K_c, M_c (or S_c, in case of model B) for each class c, resulting in a parametric estimate $\widehat{f}_c(x)$ for the class density $f_c(x)$. In the end a parametric classifier can be constructed in the usual way; a candidate object with feature vector x is assigned to the class with highest value of $p_c\widehat{f}_c(x)$, where p_c is the prior probability for class c. Alternative versions, incorporating different losses, for example, can also be constructed with the $\widehat{f}_c(x)$'s as basis.

3. Experimental results

Our aim was to use a parametric model wider than the multinormal model such that the new density function could effectively approximate multivariable distributions with both bimodal and unimodal characteristics. To test the usefulness of the linear mixture methods of Section 2 we have compared the estimated probability densities and the classification results obtained by using model A and model B with those obtained by applying the ordinary multinormal probability density model. The nonparametric k-nearest-neighbour (k-NN) classification method was included for completeness. These methods were tested on three different data sets, two from symbol recognition and one from remote sensing.

When carrying out statistical classification separate training and test sets were used (thus giving honest error rates). The prior probabilities used in classifying the uppercase, printed letters and the handwritten digits were taken to be equal for all classes. A certain image from the Thematic Mapper satellite was recorded over a region of Norway mostly covered with forest. For this image the prior probabilities were estimated by an automatic updating procedure (Hjort and Taxt, 1987).

3A. Data sets and feature measurement methods. Table 1 shows some of the characteristics of the three data sets. The sets A210 and A220 consisted of printed uppercase

letters, the sets MCENR1 and MCENR2 of handwritten digits, and the Thematic Mapper sets of pixels in a multispectral image. Figures 1a, 1b give examples of the printed and handwritten symbols, respectively. The symbol sets came from analog maps that were digitised automatically, using the automated data capture system of SysScan Ltd., Norway. The symbol training sets were labelled manually. The ground truth for the Thematic Mapper study was obtained through field survey and plotted on a map, vectorised and run through a vector-to-raster conversion algorithm. The pixels containing ground truth were then divided into a training and a test set by applying a mask selecting every second pixel line of the image.

For these experiments we used the 'grid method' features for symbols with known orientation, and for rotation invariant features the Fourier method of Zahn and Roskies (1972) (see Holbæk-Hanssen, Bråten and Taxt, 1986 for more details). For the Thematic Mapper image the feature vector of each pixel was constructed by using the value of the pixel in the six available channels (band $1, 2, 3, 4, 5, 7$) directly.

-- Figures 1a and 1b around here, please --

3B. Statistical characteristics of the data sets. The three pairs of data sets, with the chosen feature measurement methods, gave rise to increasingly complex distributions in feature space. To express this relationship quantitatively, we estimated the generalised Mahalanobis distances (see Hjort, 1986, Ch. 10, and Taxt, Ólafsdóttir and Dæhlen, 1990) between all pairs of classes in each of the three training sets. A rough rule of thumb about these is that if the distance for two specific classes is greater than ≈ 4.0, the probability of incorrect classification between them is $\approx 2.5\%$ (Hjort, 1986). We also inspected all the histograms (or nonparametric density estimates) of all feature vector components. We used the Fourier method with 3 amplitudes and 3 F-terms when studying the statistical characteristics of the symbol sets. For the satellite image we used all 6 channel images.

PRINTED LETTERS. The generalised Mahalanobis distance was always larger than 10.0, and seldom less than 20.0. The only exception was the distance between 'X' and 'O', which was 5.2. Almost all feature vector components were unimodally distributed and well modelled by both the multinormal and linear mixture models.

HANDWRITTEN DIGITS. The handwritten digits had substantially smaller interclass distances than the printed letters. Several of the interclass distances were in the range 4.0 to 7.0 and rather few were larger than 20.0. Some of the F-term feature components were clearly bimodally and sometimes even multimodally distributed. Almost all of these feature distributions were well approximated by the linear mixture models, but not by the ordinary multinormal model.

THEMATIC MAPPER IMAGE. The interclass distances were clearly smaller than those of the handwritten digits. Several of the interclass distances were in the range 1.0 to 4.0 and rather few were larger than 10.0. This meant that higher error rates had to be expected. Several of the feature component distributions were clearly bimodal, and sometimes even

multimodal. Again, most of the feature distributions were well approximated by the linear mixture models, but in many cases not by the multinormal.

3C. Significance of the size of the training set. To compare the performance of the linear mixture and multinormal based classifiers obtained by different sizes of the training set, two versions of the Fourier feature measurement method were used on the data sets of handwritten digits. The results of training on an increasing number of symbols per class, up to the maximum available number, and then classifying the whole test set using these features are summarised in Table 2. The classification rates there are the result of a particular (but random) choice of initial training sets. Other random choices gave comparable results.

In the first case only three F-terms and no amplitudes were selected. This choice was made to compare the parametric probability density models in classifier performance when the interclass distances in the feature space were low and several feature vector components bimodally or multimodally distributed. As seen from Table 2, the linear mixture model B outperformed the multinormal model even for the smallest size of the training set studied. With larger training sets both linear mixture model A and B gave better classification results than the multinormal model. The better performance of model B than model A with small training sets is explained by the smaller number of parameters that have to be estimated in model B.

In the second case 9 amplitudes and 3 F-terms were used. This choice was made to compare the probability density models in classifier performance when most interclass distances in the feature space were rather large and the number of parameters to be estimated in the probability density models high. Also, some of the components of the feature vector components were bimodally or multimodally distributed.

As seen from Table 2, the linear mixture models competed favourably with the multinormal model also in this case. The better classifier performance of the multinormal model using small training sets was not impressive. Also, the unexpected fall in classifier performance of the multinormal model with increasing number of symbols in the training set was not observed using the linear mixture models. Interestingly, this fall in the classifier performance was due to the misclassification of the symbol class '1' into symbol class '8'. This might be explained by several bimodally distributed feature vector components of class '1'. Such components were well approximated by the linear mixture models, but poorly approximated by the multinormal model.

3D. Classification with maximal training sets. As noted above, the interclass distances of the three different pairs of data sets were very different. We expected this to be reflected in the differences of the correct classification rate obtained by using the various probability density models. This was also the case. For the handwritten digits, there was a slight, but consistent improvement in the correct classification rates when bimodal components were present in the feature vector. Finally, in classifying the Thematic Mapper image, we found a substantial increase in the correct classification rates when using the linear

mixture models (Table 3).

UPPERCASE PRINTED LETTERS. For the 'grid' method we used a 3×3 grid and the sums over rows and columns, resulting in 5-dimensional feature vectors. Applying the Fourier method, the first 5 amplitudes were taken as features. Because of the comfortingly large interclass Mahalanobis distances we expected all methods to give similar classification results. This was also observed for the parametric models, but for the k-NN rule ($k = 1, 2, 5, 10$) the error rate was slightly larger for the Fourier method.

HANDWRITTEN DIGITS. For the 'grid' feature measurement method we used a 3×3 grid and the sums over the rows and the columns. Also four of the coordinates of the end nodes were included, such that the feature vector dimension became 9. Applying the Fourier method, the first 9 amplitudes and the 3 F-terms were taken as feature components. Both linear mixture models gave better classification results than the multinormal probability density model, but the improvements were small (see Table 3). The reasons for this were probably twofold. First, only a moderate number of the feature vector components had clear interclass bimodal or multimodal structure. Secondly, based on the Mahalanobis distances, almost all pairs of classes were rather well separated in feature space. This makes the approximation done by applying the normal model to bimodal or multimodal feature components less significant. The k-NN model ($k = 1, 2, 5, 10$) gave similar results to the parametric models for both feature measurement methods.

THEMATIC MAPPER IMAGE. In contrast to the two previous data sets, the separation in feature space between pairs of classes for this data set was rather small (see above). Also, many feature vector components were bimodally or multimodally distributed. We observed a substantial improvement by applying the linear mixture models compared to the multinormal model. This was the case for both the noncontextual and contextual method we applied. The linear mixture model A gave close to 1% better than the corresponding results with the linear mixture model B. It is worth noting that Owen's (1984) contextual method (and other contextual methods) gave about 5% improvement in correct classification when compared to the noncontextual method. The k-NN rule gave results slightly superior to ($k = 10$) or comparable to ($k = 5$) the multinormal model in all cases.

4. Summary and conclusions

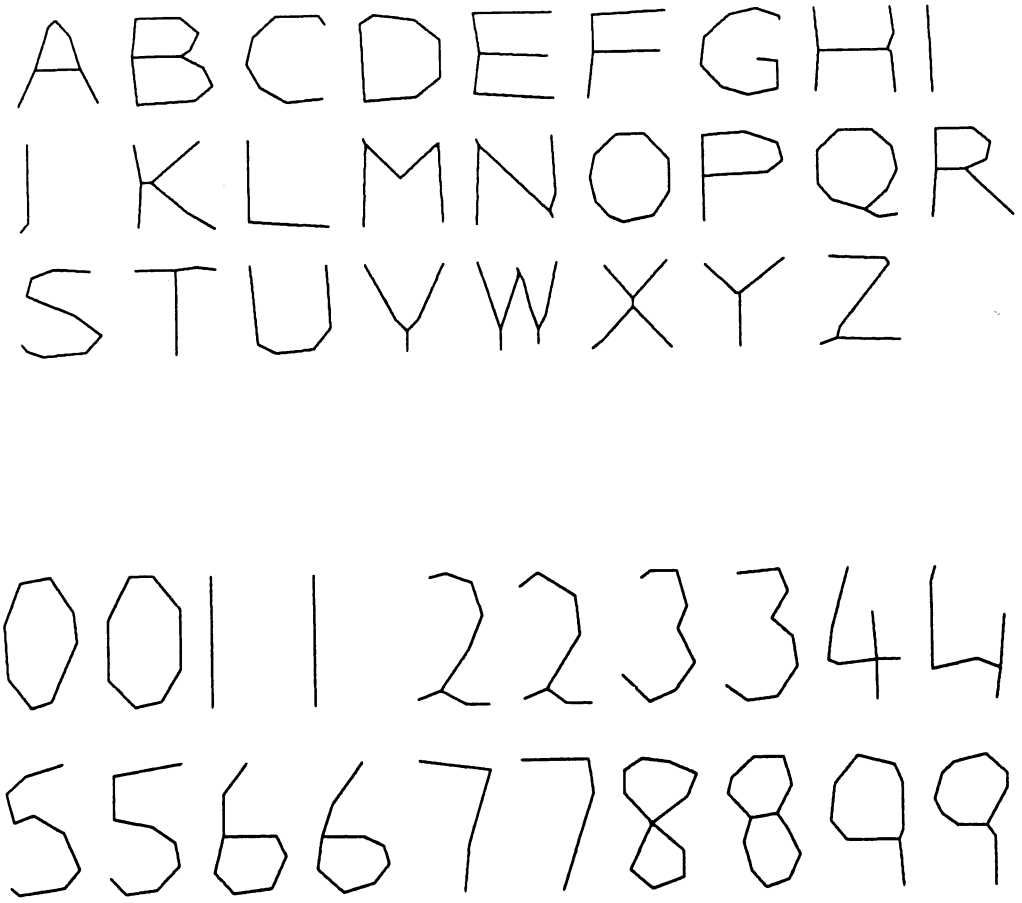
We have developed a method based on an EM algorithm to estimate the parameters of a probability density model consisting of a linear mixture of two multinormal distributions. The empirical results demonstrate that these linear mixture methods can be superior to the multinormal method and the k-NN rule in statistical classification of data sets with small interclass distances. However, the training set has to be of a reasonable size compared to the number of parameters to be estimated. For small training sets or larger interclass distances the performance of the linear mixture models is comparable to the multinormal model.

The choice between a linear mixture model with a common (B) or two separate covariance matrices (A) is strongly data dependent. Model choice criteria like Akaike's or

Schwarz' could be used (see Hjort, 1986, Ch. 7). If the training set is large, method A is expected to perform better than method B because of the larger flexibility of model A in fitting bimodal feature vector components with different widths of the individual peaks. One might also try out models in between, i.e. with some common covariance structure, or with different weights for different sets of components.

In conclusion, we have shown that it is possible to obtain good estimates of the parameters in linear mixture models of two multinormal distributions. This allows a significant improvement in statistical classification of some data sets with bimodally distributed feature vector components.

Acknowledgement. This work has been supported in part by SysScan a.s., Norway, and the Royal Norwegian Council for Scientific and Industrial Research.



FIGURES 1A AND 1B. (a) shows handwritten letters (after pre-processing) from the training set A210. (b) diplays handwritten numbers (after pre-processing) from the the training set MCENR1.

data set	*size*	*number of classes*	*feature measure. method*	*feature vector dim.*
A210 (train)	3874	26	Grid, Fourier	5, 5
A220 (test)	3890	26	Grid, Fourier	5, 5
MCENR1 (train)	3987	10	Grid, Fourier	9, 12
MCENR2 (test)	5093	10	Grid, Fourier	9, 12
The.Map.(train)	2046	9	Identity	6
The. Map (test)	2096	9	Identity	6

Table 1. *Properties of data sets and feature measurement.* The size is the number of labelled symbols or the number of pixels in the sets.

	3D feature vector			12D feature vector		
symbols per class	*multi-normal*	*mixture model A*	*mixture model B*	*multi-normal*	*mixture model A*	*mixture model B*
20	67.3 ± 1.3	68.7 ± 2.9	68.5 ± 3.6	87.0 ± 1.5	-	86.1 ± 1.2
50	68.8 ± 0.3	74.6 ± 1.3	72.5 ± 1.2	93.9 ± 0.6	90.0 ± 1.6	93.3 ± 0.9
100	69.0 ± 0.4	76.6 ± 0.9	74.3 ± 0.9	95.6 ± 0.4	95.2 ± 0.0	95.8 ± 0.5
400	71.3	74.7	73.2	98.7	98.8	98.8

Table 2. *Size of training set and classifier performance.* The percent correct classification (mean±S.D.) of the handwritten digits as a function of the size of the training set.

	Grid method				Fourier method			
data set	*multi-normal*	*10-NN*	*mixture model A*	*mixture model B*	*multi-normal*	*10-NN*	*mixture model A*	*mixture model B*
printed letters	99.9 6	99.9 5	99.8 8	99.9 6	99.9 6	98.6 54	99.8 8	99.8 7
handwritten digits	99.2 39	98.8 60	99.4 33	99.3 36	98.7 68	99.0 50	98.8 62	98.8 61
	Noncontextual method				Owen's contextual method			
Thematic Mapper	78.6 434	79.3 420	80.7 392	79.3 420	82.6 353	83.9 326	85.6 292	84.6 312

Table 3. *Correct classification rates and number of errors.* The first line of each data set is the % correct classification, the second line the number of errors.

References

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* Academic Press, New York.

Bezdek, J.C., Hathaway, R.J., and Huggins, V.J. (1985). Parametric estimation for normal mixtures. *Pattern Recognition Letters* **3**, 79–84.

Hjort, N.L. (1986). *Theory of Statistical Symbol Recognition.* Research monograph, Norwegian Computing Centre.

Hjort, N.L. (1991). On some tests for presence of clusters. Technical report, Norwegian Computing Centre, Oslo.

Hjort, N.L. and Taxt, T. (1987). Automatic training in statistical pattern recognition. In *Image analysis and processing II*, eds. Cantoni, Di Gesu and Levialdi. Plenum Press, New York.

Holbæk-Hanssen, E., Bråten, K.H., and Taxt, T. (1986). A general software system for supervised statistical classification of symbols. *ICPR Proc. of the Eighth Int. Conf. on Pattern Recognition*, Paris, 144–149.

Hosmer, D.W. (1978). A use of mixtures of two normal distributions in a classification problem. *J. Statist. Comput. Simul.* **6**, 137–148.

Lachenbruch, P.A. and Brofitt, B. (1980). On classifying observations when one population is a mixture of normals. *Biom. J.* **22**, 295–301.

Owen, A. (1984). A neighbourhood-based classifier for Landsat data. *Canadian J. Statist.* **12**, 191–200.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

Taxt, T., Eikvil, L., and Hjort, N.L. (1989). Statistical classification using mixtures of multinormal densities. In *Proceedings from the ESA conference EXPERTSYS-89.*, ed. Liebowitz. IITT-International, Gournay-sur-Marne, France.

Taxt, T., Ólafsdóttir, J.B., and Dæhlen, M. (1990). Recognition of handwritten symbols. *Pattern recognition* **23**, 1155–1166.

Zahn, C.T. and Roskies, R.Z. (1972). Fourier descriptors for plane closed curves. *IEEE Trans. on Computers* C-21, No.3, 269–281.