

# Model-based mixture discriminant analysis – an experimental study

Zohar Halbe and Mayer Aladjem

Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev  
P.O.Box 653, Beer-Sheva, 84105, Israel

## Abstract

The subject of this paper is an experimental study of a discriminant analysis (DA) based on Gaussian mixture estimation of the class-conditional densities. Five parameterizations of the covariance matrixes of the Gaussian components are studied. Recommendation for selection of the suitable parameterization of the covariance matrixes is given.

Keywords: Discriminant analysis, Gaussian mixture model, Density estimation, Model selection.

## 1. Introduction

*Discriminant analysis* (DA) is a powerful technique for classifying observations into known pre-existing classes. In the Bayesian decision framework [1] a common assumption is that the observed  $d$ -dimensional patterns  $\mathbf{x}$  ( $\mathbf{x} \in \mathbb{R}^d$ ) are characterized by the *class conditional density*  $f_c(\mathbf{x})$ , for each class  $c=1, 2, \dots, C$ . Let  $P_c$  denotes a prior probability of the class  $c$ . According to Bayes theorem the posterior probability that an arbitrary observation  $\mathbf{x}$  belongs to class  $c$  is  $P(\mathbf{x} \in \text{Class } c | \mathbf{x}) = P_c f_c(\mathbf{x}) / \sum_{j=1}^C P_j f_j(\mathbf{x})$ . The classification rule for allocating  $\mathbf{x}$  to the class  $c$  having the highest posterior probability  $P(\mathbf{x} \in \text{Class } c | \mathbf{x})$ , minimizes the expected misclassification rate [1]. The latter rule is named the Bayes classification rule.

In this paper we assume that  $f_c(\mathbf{x})$  is a mixture of  $M_c$  normal (Gaussian) densities

$$f_c(\mathbf{x}) = \sum_{j=1}^{M_c} \pi_{cj} N(\mathbf{x} | \boldsymbol{\mu}_{cj}, \boldsymbol{\Sigma}_{cj}). \quad (1)$$

Here  $\pi_{cj}$  are *mixing coefficients*, which are non-negative and sum to one.  $N(\mathbf{x} | \boldsymbol{\mu}_{cj}, \boldsymbol{\Sigma}_{cj})$  denotes multivariate normal density with mean vector  $\boldsymbol{\mu}_{cj}$  and covariance matrix  $\boldsymbol{\Sigma}_{cj}$ . Usually the covariance matrixes  $\boldsymbol{\Sigma}_{cj}$  are taken to be *full* (unrestricted) or *diagonal* or *spherical*. The fitting of the parameters  $\pi_{cj}$ ,  $\boldsymbol{\mu}_{cj}$  and  $\boldsymbol{\Sigma}_{cj}$  is carried out by maximizing the likelihood of the parameters to the training data  $\{\mathbf{x}_{c1}, \mathbf{x}_{c2}, \dots, \mathbf{x}_{cn_c}\}$  of  $n_c$  observations from class  $c$ .

The subject of this paper is an experimental study of a DA based on an extended parameterization of  $\boldsymbol{\Sigma}_{cj}$ , proposed in [3] and named *model-based DA* [2]. In addition, we modify the model selection

procedure proposed in [2], which overcomes the setting of the maximal number of components for the mixture (1). To the best of our knowledge this is the first extensive experimental study of the model-based DA. The paper is organized as follows: in Section 2, we describe the model-based DA, experimental results are provided in Section 3 and conclusions in Section 4.

## 2. Model-based DA

Following [2, 3] the parameterizations of the covariance matrixes  $\Sigma_{cj}$  are taken to be:  $\Sigma_{cj}=\lambda_{cj}\mathbf{I}$ ;  $\mathbf{I}$  denotes the identity matrix and  $\lambda_{cj}$  are adjustable parameters;  $\Sigma_{cj}=\lambda_{cj}\mathbf{B}_c$ ;  $\Sigma_{cj}=\lambda_{cj}\mathbf{B}_{cj}$ , where  $\mathbf{B}_c$  and  $\mathbf{B}_{cj}$  are adjustable diagonal matrixes having positive diagonal elements;  $\Sigma_{cj}=\lambda_{cj}\mathbf{D}_c$ ; and  $\Sigma_{cj}=\mathbf{D}_{cj}$ , where  $\mathbf{D}_c$  and  $\mathbf{D}_{cj}$  are adjustable positive definite symmetric matrixes.  $\mathbf{D}_c$  is restricted to have unit determinant ( $|\mathbf{D}_c|=1$ ). Note, that  $\Sigma_{cj}=\lambda_{cj}\mathbf{I}$ ,  $\lambda_{cj}\mathbf{B}_{cj}$  and  $\mathbf{D}_{cj}$  are the spherical, diagonal and full covariance matrixes, respectively, used in the conventional mixture models (1). The other parameterizations  $\Sigma_{cj}=\lambda_{cj}\mathbf{B}_c$ ,  $\lambda_{cj}\mathbf{D}_c$  define  $f_c(\mathbf{x})$  (1) having the same matrixes  $\mathbf{B}_c$  and  $\mathbf{D}_c$  for all mixture components of  $f_c(\mathbf{x})$ . The goal is to reduce the number  $v_c$  of the adjustable parameters of  $f_c(\mathbf{x})$  while ensuring flexibility of the mixture model by adjusting the parameters  $\lambda_{cj}$  for each component. Table 1 gives the expressions for  $v_c$ , for the parameterizations of  $\Sigma_{cj}$  used in this paper.

Table 1: Expressions for  $\Sigma_{cj}$  parameterization and the number  $v_c$  of the adjustable parameters of  $f_c(\mathbf{x})$ .

$\Sigma_{cj}$	$\lambda_{cj}$	$\mathbf{B}_c$ (or $\mathbf{B}_{cj}$ )	$\mathbf{D}_c$ (or $\mathbf{D}_{cj}$ )	$v_c$
$\lambda_{cj}\mathbf{I}$	$\text{tr}(\mathbf{W}_{cj})/dn_{cj}$	-	-	$d(M_c+1)+M_c-1$
$\lambda_{cj}\mathbf{B}_c$	$\text{tr}(\mathbf{W}_{cj}\mathbf{B}_c^{-1})/dn_{cj}$	$\text{diag}(\sum_j \lambda_{cj}^{-1}\mathbf{W}_{cj})/ \text{diag}(\sum_j \lambda_{cj}^{-1}\mathbf{W}_{cj}) ^{1/d}$	-	$d(M_c+1)+2(M_c-1)$
$\lambda_{cj}\mathbf{B}_{cj}$	$\text{tr}(\mathbf{W}_{cj}\mathbf{B}_{cj}^{-1})/dn_{cj}$	$\text{diag}(\mathbf{W}_{cj})/ \text{diag}(\mathbf{W}_{cj}) ^{1/d}$	-	$2M_c d+M_c-1$
$\lambda_{cj}\mathbf{D}_c$	$\text{tr}(\mathbf{W}_{cj}\mathbf{D}_c^{-1})/dn_{cj}$	-	$\sum_j \lambda_{cj}^{-1}\mathbf{W}_{cj}/ \sum_j \lambda_{cj}^{-1}\mathbf{W}_{cj} ^{1/d}$	$2(M_c-1)+M_c d+d(d+1)/2$
$\mathbf{D}_{cj}$	1	-	$\mathbf{W}_{cj}/n_{cj}$	$M_c d+M_c-1+M_c d(d+1)/2$

For a predefined number of components  $M_c$  and parameterization of  $\Sigma_{cj}$  (described above), the parameters  $\mu_{cj}$ ,  $\Sigma_{cj}$  and  $\pi_{cj}$  are determined by an *expectation-maximization* (EM) algorithm proposed in [3]. The EM algorithm is initialized by the k-means clustering algorithm [1]. Using the latter algorithm the training data  $\{\mathbf{x}_{c1}, \mathbf{x}_{c2}, \dots, \mathbf{x}_{cn_c}\}$  is partitioned into  $M_c$  groups  $G_{cj}$ ,  $j=1, 2, \dots, M_c$ . An indicator vector

$\mathbf{z}_{ci} = (z_{i1}^{(c)}, z_{i2}^{(c)}, \dots, z_{iM_c}^{(c)})$  is associated with each observation  $\mathbf{x}_{ci}$  from class  $c$ . For  $\mathbf{x}_{ci} \in G_{cj}$ ,  $z_{ij}^{(c)} = 1$ , otherwise  $z_{ij}^{(c)} = 0$ , for  $i=1, \dots, n_c$  and  $j=1, \dots, M_c$ . The EM algorithm alternates between two steps, an *expectation step* (E-step) and a *maximization step* (M-step), until convergence criterion is satisfied. Using the initial  $\mathbf{z}_{ci}$  and the initial  $\lambda_{cj}=1$ , for  $j=1, \dots, M_c$ , the following E- and M-steps are cycled:

$$\text{M-step: } n_{cj} \leftarrow \sum_{i=1}^{n_c} z_{ij}^{(c)}, \quad \boldsymbol{\mu}_{cj} \leftarrow \frac{\sum_{i=1}^{n_c} z_{ij}^{(c)} \mathbf{x}_{ci}}{n_{cj}}, \quad \pi_{cj} = \frac{n_{cj}}{n_c}, \quad \mathbf{W}_{cj} = \sum_{i=1}^{n_c} z_{ij}^{(c)} (\mathbf{x}_{ci} - \boldsymbol{\mu}_{cj})(\mathbf{x}_{ci} - \boldsymbol{\mu}_{cj})^T$$

Calculate  $\boldsymbol{\Sigma}_{cj}$  using the expressions in Table 1.

$$\text{E-step: } z_{ij}^{(c)} \leftarrow \frac{\pi_{cj} N(\mathbf{x}_{ci} | \boldsymbol{\mu}_{cj}, \boldsymbol{\Sigma}_{cj})}{\sum_{j=1}^{M_c} \pi_{cj} N_j(\mathbf{x}_{ci} | \boldsymbol{\mu}_{cj}, \boldsymbol{\Sigma}_{cj})}.$$

In the model-based DA [2] each combination of the number  $M_c$  of the components of  $f_c(\mathbf{x})$  (1) and the parameterization of  $\boldsymbol{\Sigma}_{cj}$  corresponds to a certain Gaussian mixture model. In [2] the *Bayesian information criterion* (BIC) is used for model selection. The  $BIC(M_c)$  for  $f_c(\mathbf{x})$  (1) having  $M_c$  components is:

$$BIC(M_c) = -2 \sum_{i=1}^{n_c} \log f_c(\mathbf{x}_{ci}) + v_c \log(n_c), \quad (2)$$

where  $f_c(\mathbf{x}_{ci})$  is the mixture model (1) fitted to the training data by the EM algorithm. Using  $BIC(M_c)$  (2) the number  $M_c$  and the parameterization of  $\boldsymbol{\Sigma}_{cj}$  are set by the following procedure. For each parameterization of  $\boldsymbol{\Sigma}_{cj}$ :

- (1) Set  $M_c=0$ ,  $BIC(0)=\infty$ .
- (2) Update  $M_c=M_c+1$ . Apply the EM algorithm for fitting the  $f_c(\mathbf{x})$ . Then compute  $BIC(M_c)$  (2).
- (3) If  $BIC(M_c) > BIC(M_c-1)$  then set  $M_c = M_c-1$  else repeat steps (2)-(3).

Finally we select the model (parameterization of  $\boldsymbol{\Sigma}_{cj}$  and the number  $M_c$  of the components) corresponding to the minimal value of BIC.

This procedure is a modification of the original proposal [2]. It overcomes the problem in setting the maximal number  $M_{c \max}$  of components required in [2] and reduces the computational complexity.

### 3. Experiments

In this section we investigated the performance of the model-based DA. For comparison we ran the *probabilistic principal component analysis* (PPCA) mixture density estimation [4], which is a popular latent variable method in pattern recognition literature. We carried out experiments on various data sets, artificial and real-world data, from the UCI Machine Learning Repository at <http://www.ics.uci.edu/~mllearn/MLRepository.html> and Gunnar Ratsch at <http://www.first.gmd.de/~raetsch/data/banana.txt>. In addition to those benchmark data sets we composed a data set named MODIFIED LETTER. It merges the classes (26 letters) of the benchmark data set LETTER into two groups. We set letters O, U, P, S, X, Z, E, B, F, T, W, A, Q to be Group 1 and the letters H, D, N, C, R, G, K, Y, V, M, I, J, L to be Group 2, using cluster analysis of the letters means. The goal was to set complicated (highly overlapped) groups. In our experiments with MODIFIED LETTER we set a training sample size of 300 observations, selected from the original LETTER data randomly. Table 2 shows the characteristics of the data sets: C is the number of the classes, S is the averaged number of the observations ( $S = \sum_c n_c / C$ ) into the classes, p is the original dimension of the observations and d is the reduced dimension after preprocessing by principal component analysis [1]. We indicate by \* the data sets having small sample sizes ( $S/d \leq 10$ ).

Table 2: Data sets characteristics.

DATA	C	p	d	S/d
SONAR*	2	60	29	2.86
GLASS*	6	9	6	5.83
IONOSPHERE*	2	34	30	5.83
VOWEL*	11	13	9	9.88
MODIFIED LETTER*	2	16	15	10
IRIS	3	4	3	16.66
WINE	3	13	12	19.66
LIVER DISORDERS	2	7	5	27.6
GAUSSIAN	2	8	7	35.71
WAVEFORM-NOISE	3	40	40	41.65
LETTER	26	16	15	51.26
WAVEFORM	3	21	21	79.33
BANANA	2	2	2	1250
BANANA-NOISE	2	2	2	1250

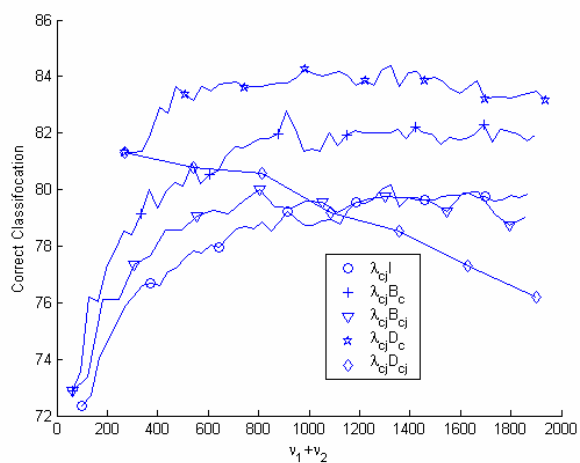


Figure 1: Correct classification rate versus number of components for MODIFIED LETTER data set.

Each experiment consisted of 10 runs of the following procedure. We randomly drew a data set with replacement from the original data and split the data into five roughly equal sized parts. Then we fitted the parameters of  $f_c(\mathbf{x})$  (1) and selected the best model ( $M_c$  and parameterization of  $\Sigma_{c_j}$ ) using four parts of the data and calculated the correct classification rate, allocating the observations from the left part by the Bayes classification rule (Section 1). We carried out five replications using a different part for classification each time. Finally, we averaged the classification rate over  $10 \times 5$  random runs. (When a covariance matrix  $\Sigma_{c_j}$  happened to be singular, the zero eigenvalues were replaced by a small number just large enough to permit numerically stable inversion).

In Table 3, we report the obtained averaged percentage of the correct classification rate along with the standard errors. In parentheses we report the averaged number of the model parameters ( $\sum_c \nu_c$ ). The bold scores indicate the highest percentage of correct classification rates.

Table 3: Averaged percentage of correct classification rates.

DATA	DA with $f_c(\mathbf{x})$ having different parameterizations of $\Sigma_{c_j}$					PPCA
	$\lambda_{c_j} \mathbf{I}$	$\lambda_{c_j} \mathbf{B}_c$	$\lambda_{c_j} \mathbf{B}_{c_j}$	$\lambda_{c_j} \mathbf{D}_c$	$\mathbf{D}_{c_j}$	
SONAR*	71.9±0.89(315)	74.8±0.85(197)	74.7±0.92(232)	<b>81.3±0.64(996)</b>	80.9±0.62(928)	79.8±0.73(456)
GLASS*	57.9±0.98(147)	<b>62.2±0.71(175)</b>	61.5±0.82(194)	61.0±0.82(279)	61.0±0.82(279)	49.0±0.92(90)
IONOSPHERE*	91.4±0.42(375)	<b>92.5±0.40(404)</b>	92.5±0.39(471)	91.9±0.40(1106)	83.1±0.69(1496)	81.7±0.52(126)
VOWEL*	85.3±0.40(760)	86.2±0.40(808)	87.2±0.40(1084)	<b>89.9±0.28(1083)</b>	87.7±0.35(1868)	85.4±0.28(680)
MODIFIED LETTER*	79.4±0.73(203)	81.6±0.52(229)	79.8±0.52(242)	<b>82.0±0.56(330)</b>	81.7±0.50(273)	81.0±0.61(258)
IRIS	96.1±0.46(40)	95.5±0.53(36)	94.8±0.50(35)	96.9±0.41(27)	96.9±0.41(27)	<b>97.4±0.35(28)</b>
WINE	96.0±0.36(137)	96.8±0.36(137)	96.6±0.41(105)	98.3±0.30(279)	<b>98.7±0.28(270)</b>	98.4±0.28(137)
LIVER DISORDERS	66.1±0.50(71)	65.1±0.49(63)	64.1±0.63(70)	67.4±0.65(66)	<b>67.9±0.69(95)</b>	60.7±0.71(56)
GAUSSIAN	69.8±0.57(100)	71.6±0.64(84)	69.8±0.62(86)	81.0±0.48(73)	<b>84.8±0.35(105)</b>	73.4±0.57(103)
WAVEFORM-NOISE	<b>86.7±0.14(446)</b>	86.5±0.13(550)	86.5±0.13(807)	84.0±0.14(2580)	84.0±0.14(2580)	86.0±0.14(249)
LETTER	82.7±0.12(5450)	89.0±0.10(5614)	90.0±0.07(8133)	91.7±0.07(6987)	<b>94.73±0.07(13384)</b>	84.0±0.07(3380)
WAVEFORM	<b>86.5±0.10(297)</b>	86.4±0.09(355)	86.3±0.11(427)	85.1±0.13(787)	85.1±0.12(756)	86.0±0.12(135)
BANANA	92.9±0.12(65)	94.2±0.10(74)	94.2±0.10(72)	94.3±0.10(75)	<b>95.1±0.06(61)</b>	95.1±0.07(71)
BANANA-NOISE	75.3±0.14(37)	75.1±0.15(36)	75.4±0.15(36)	75.4±0.14(46)	<b>75.8±0.16(37)</b>	75.8±0.16(45)

For the MODIFIED LETTER data set we ran an additional experiment. We trained the conditional densities (1) using the selected 300 training observations and calculated the correct classification rate using the Bayes classification rule for the rest of the 19700 LETTER observations.

In Fig.1 we show the averaged classification rates versus the number  $\nu_1 + \nu_2$  of the parameters of the

model. We constructed the graphs using the same number of components for each class ( $M_1=M_2=M$ ) varying  $M=1, 2, \dots, 20$ .

Observing the results in Table 3 and Fig. 1 we conclude that the model  $\lambda_{c_j}\mathbf{D}_c$  outperforms others models for the data sets SONAR, VOWEL and MODIFIED LETTER having a small sample size ( $S/d \leq 10$ ) and the model  $\lambda_{c_j}\mathbf{B}_c$  outperforms the model  $\lambda_{c_j}\mathbf{B}_{c_j}$  for most of the data sets. Consequently it seems that the models  $\Sigma_{c_j}=\lambda_{c_j}\mathbf{I}$ ,  $\lambda_{c_j}\mathbf{B}_c$  and  $\Sigma_{c_j}=\lambda_{c_j}\mathbf{D}_c$  cover a wide range of data structures and could be considered as the basic set of models for the model-based DA. Looking at Table 3 for the MODIFIED LETTER dataset we observe that using BIC we selected models  $\lambda_{c_j}\mathbf{B}_c$  and  $\lambda_{c_j}\mathbf{D}_c$  having  $v_1+v_2 = 229-330$ , respectively. Observing the results in Fig.1 we conclude that for those models ( $\lambda_{c_j}\mathbf{B}_c$ ,  $\lambda_{c_j}\mathbf{D}_c$ ) the best generalization performance (highest classification rate for 19700 test observations) is for  $v_1+v_2 > 800$ . Consequently, the model selection using BIC underestimates the complexity of  $f_c(\mathbf{x})$  used for Bayes classification. An improvement of the model selection is needed. The latter is a subject of our future research.

#### 4. Conclusions

In this paper, we have investigated the performance of the model-based DA [2, 3] on various data sets, artificial and real-world data. The results obtained in the experimental study show that  $\Sigma_{c_j}=\lambda_{c_j}\mathbf{I}$ ,  $\lambda_{c_j}\mathbf{B}_c$  and  $\lambda_{c_j}\mathbf{D}_c$  seems to be a universal set of models that can be recommended for practical applications in small sample size data sets. In addition, we proposed a modification of the model selection procedure proposed in [2], which overcomes the setting of the maximal number of components in the mixture (1) and reduces the computation complexity of the model selection.

#### References

- [1] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer press, Canada, 2001.
- [2] C. Fraley, A.E. Raftery, Model-Based Clustering, Discriminant Analysis, and Density Estimation, Journal of the American Statistical Association, Vol. 97, pp. 611-631, Jun 2002.
- [3] G. Celeux, G. Govaert, Gaussian parsimonious clustering models, Pattern Recognition, Vol. 28, No. 5, pp. 781-793, 1995.
- [4] M.E. Tipping, C.M. Bishop, Mixture of Probabilistic Principal Component Analyzers, Neural Computation, Vol. 11, pp. 443-482, February 1999.

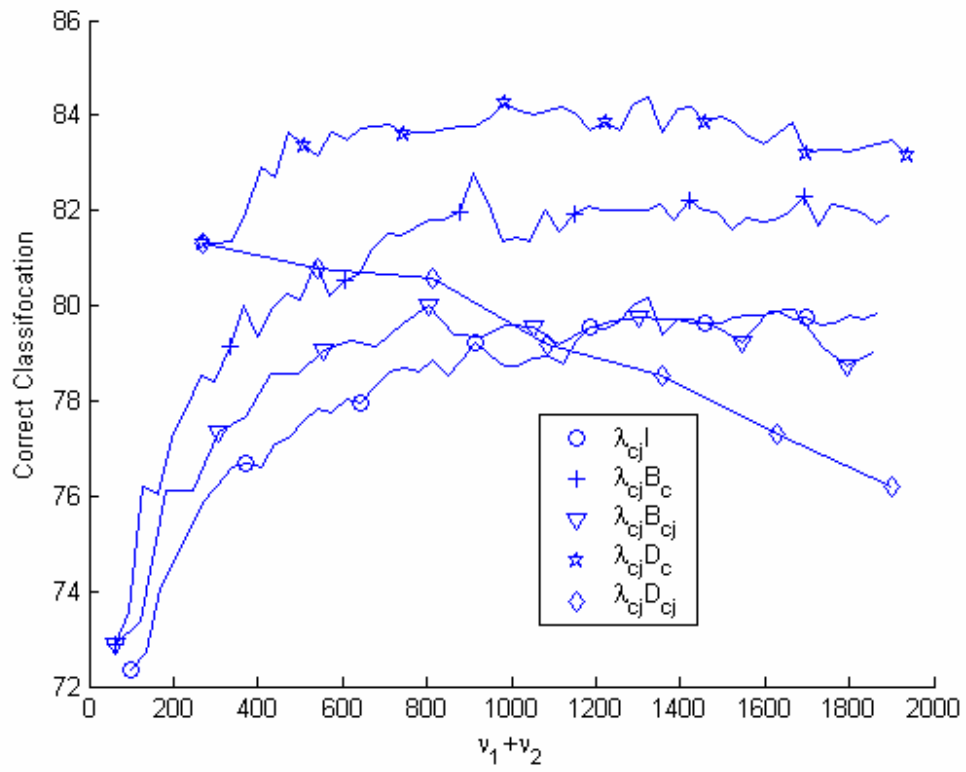


Figure 1: Correct classification rate versus number of components for MODIFIED LETTER data set.