

Regularized Mixture Discriminant Analysis

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 – In this paper we seek a Gaussian mixture model (GMM) of the class-conditional densities for plug-in Bayes classification. We propose a method for setting the number of the components and the covariance matrices of the class-conditional GMMs. It compromises between simplicity of the model selection based on the Bayesian information criterion (BIC) and the high accuracy of the model selection based on the cross-validation (CV) estimate of the correct classification rate. We apply an idea of Friedman (1989) to shrink a predefined covariance matrix to a parameterization with substantially reduced degrees of freedom (reduced number of the adjustable parameters). Our method differs from the original Friedman's method by the meaning of the shrinkage. We operate on matrices computed for a certain class while the Friedman's method shrinks matrices from different classes. We compare our method with the conventional methods for setting the GMMs based on the BIC and CV. The experimental results show that our method has the potential to produce parameterizations of the covariance matrices of the GMMs which are better than the parameterizations used in other methods. We observed significant enlargement of the correct classification rates for our method with respect to the other methods which is more pronounced as the training sample size decreases. The latter implies that our method could be an attractive choice for applications based on a small number of training observations.

Key Words - Gaussian mixture models, Model selection, Bayesian information criterion, Classification, Regularized discriminant analysis.

1. Introduction

The purpose of *discriminant analysis* (DA) is to classify observations into known preexisting classes $c=1, 2, \dots, C$. An observation is assumed to be a member of one (and only one) class and an error is incurred if it is assigned to a different one. In a Bayesian decision framework (Duda, Hart and Stork 2001; Hastie, Tibshirani and Friedman 2001; McLachlan 2004) 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 $P(c|\mathbf{x})$ that an arbitrary observation \mathbf{x} belongs to class c is

$$P(c|\mathbf{x}) = \frac{P_c f_c(\mathbf{x})}{\sum_{j=1}^C P_j f_j(\mathbf{x})}. \quad (1)$$

The classification rule for allocating \mathbf{x} to the class c having the highest posterior probability (1) minimizes the expected misclassification rate (Duda et al. 2001; Hastie et al. 2001; McLachlan 2004). The latter rule is called the Bayes classification rule.

In this paper we study a plug-in Bayes classification rule (McLachlan 2004) assuming a *Gaussian mixture model* (GMM) for $f_c(\mathbf{x})$ (Titterington, Smith and Makov 1985; McLachlan and Basford 1998; McLachlan and Peel 2000)

$$f_c(\mathbf{x}) = \sum_{j=1}^{M_c} \pi_{cj} \Phi_{\Sigma_{cj}}(\mathbf{x} - \boldsymbol{\mu}_{cj}). \quad (2)$$

Here, M_c denotes the number of the Gaussian components and π_{cj} are mixing coefficients, which are non-negative and sum to one. $\Phi_{\Sigma_{cj}}(\mathbf{x} - \boldsymbol{\mu}_{cj})$ denotes multivariate normal density with mean vector $\boldsymbol{\mu}_{cj}$ and covariance matrix Σ_{cj} . The fitting of the parameters π_{cj} , $\boldsymbol{\mu}_{cj}$ and Σ_{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 . Usually an *expectation maximization* (EM) algorithm (Dempster, Laird and Rubin 1977; McLachlan 1997) is applied. DA using $f_c(\mathbf{x})$ (2) for computing $P(c|\mathbf{x})$ (1) is called *mixture discriminant analysis* (MDA) (Hastie et al. 2001).

The GMM (2) is a universal approximator (Titterton et al. 1985; McLachlan and Basford 1998), i.e., we can approximate any continuous density to arbitrary accuracy, providing the model has a sufficiently large M_c . The GMM (2) is widely applied due to its easy of interpretation by viewing each fitted Gaussian component as a distinct cluster in the data. The clusters are centered at the means $\boldsymbol{\mu}_{cj}$ and have geometric features (shape, volume, orientation) determined by the covariance matrices $\boldsymbol{\Sigma}_{cj}$.

The problem of determining the number M_c and the parameterization of $\boldsymbol{\Sigma}_{cj}$ is known as *model selection* (MS). Usually several models are considered and an appropriate one is chosen using some criterion (Fraley and Raftery 1998, 2002; Biernacki and Govaert 1999; Biernacki, Celeux and Govaert 2000; Hastie et al. 2001; McLachlan 2004).

In the conventional GMMs (Titterton et al. 1985; Bishop 1995; McLachlan 2004) the covariance matrices $\boldsymbol{\Sigma}_{cj}$ are taken to be unrestricted (full) ($\boldsymbol{\Sigma}_{cj}=\mathbf{F}_{cj}$), diagonal ($\boldsymbol{\Sigma}_{cj}=\mathbf{D}_{cj}$) or spherical ($\boldsymbol{\Sigma}_{cj}=\lambda_{cj}\mathbf{I}$). Here, \mathbf{F}_{cj} denotes adjustable positive definite symmetric matrix, \mathbf{D}_{cj} is adjustable diagonal matrix having positive diagonal elements and λ_{cj} is a positive multiplier of the identity matrix \mathbf{I} . Fraley and Raftery (2002) have proposed and studied a MDA based on an extension of the conventional parameterizations of $\boldsymbol{\Sigma}_{cj}$. These parameterizations are based on a geometrical interpretation of the Gaussian clusters using eigenvalue decomposition of the $\boldsymbol{\Sigma}_{cj}$ (Banfield and Raftery 1993; Bensmail and Celeux 1996; Celeux and Govaert 1995). In our previous study (Halbe and Aladjem 2005) we have shown experimentally that only two extra parameterizations

$\Sigma_{c_j} = \lambda_{c_j} \mathbf{D}_c$ and $\lambda_{c_j} \mathbf{F}_c$, are enough for wide broad applications. These parameterizations define $f_c(\mathbf{x})$ (2) having the same component matrices, diagonal \mathbf{D}_c or unrestricted (full) \mathbf{F}_c . For these parameterizations the number of the adjustable parameters of $f_c(\mathbf{x})$ (2) has been reduced significantly (compared to the conventional diagonal and full GMMs) while ensuring flexibility of the mixture model by adjusting the parameters λ_{c_j} for each component. For the parameterizations $\Sigma_{c_j} = \lambda_{c_j} \mathbf{D}_c$ or $\lambda_{c_j} \mathbf{F}_c$ we have to compute the common component matrix \mathbf{D}_c or \mathbf{F}_c for the GMM (2), while in the conventional GMMs we have to compute M_c different matrices \mathbf{D}_{c_j} or \mathbf{F}_{c_j} , for $j=1, 2, \dots, M_c$.

It is known (Hastie et al. 2001; McLachlan 2004) that the best model selection criterion for the classification problems is the correct classification rate. Usually, the *cross-validation* (CV) estimate (Stone 1974; Hastie et al. 2001) of the correct classification rate is used for this purpose. Unfortunately, the direct application of this criterion for MDA is not practical due to the large number of GMMs (2) which need to be considered. For example, given a three class problem ($C=3$), using five parameterizations of Σ_{c_j} and a maximum number of components $M_{c \max}=15$ we have to carry out CV computations for $(5 \times 15)^3$ models. Fraley and Raftery (2002) proposed to select the parameterization of Σ_{c_j} and the number M_c for each class separately by the *Bayesian information criterion* (BIC) (Schwartz 1979) in order to reduce the computation complexity. In this case, $3 \times (5 \times 15)$ models have to be considered. Moreover, the computation of BIC is simpler than the CV estimates of the correct classification rate. Biernacki and Govaret (1999) compared experimentally variety model selection criteria for GMM with parameterizations based on the eigenvalue decomposition of Σ_{c_j} . This comparative study shows that BIC gives satisfactory results for discriminant analysis.

In this paper we propose a method which compromises between simplicity of the method of Fraley and Raftery (2002) based on BIC model selection and the high accuracy of the model selection based on the CV estimates of the correct classification rate. We carry out a two-step model selection. First, for predefined diagonal parameterization $\Sigma_{c_j} = \lambda_{c_j} \mathbf{D}_c$ the number M_c is set using BIC for each class separately. Then we shrink the predefined parameterization $\lambda_{c_j} \mathbf{D}_c$ to more complicated parameterizations of the component covariance matrices Σ_{c_j} using an idea of Friedman (1989), for regularization of Σ_{c_j} , which maximizes the correct classification rate. A comparative study of our method with the method of Fraley and Raftery (2002) and CV methods shows that our method provides a significant improvement of the classification performance, which is more pronounced as the training sample size decreases.

The paper is organized as follows. In Section 2, we describe the parameter estimation and model selection of the class-conditional GMMs (2) following the paper of Fraley and Raftery (2002). In Section 3 we present our method. We discuss the relation of our method to the method of Friedman (1989) in Section 4. The experimental results are provided in Section 5 and conclusions in Section 6.

2. EM parameter estimation and BIC model selection for the class-conditional GMMs

Based on our previous experimental study (Halbe and Aladjem 2005) we consider class-conditional GMMs (2) with $\Sigma_{c_j} = \lambda_{c_j} \mathbf{I}$, $\lambda_{c_j} \mathbf{D}_c$, \mathbf{D}_{c_j} , $\lambda_{c_j} \mathbf{F}_c$ and \mathbf{F}_{c_j} . As we have mentioned previously (Section 1) the latter parameterizations of Σ_{c_j} have been found to be suitable for broad applications.

2.1 EM algorithm

Following the paper of Fraley and Raftery (2002) we describe the estimation of the parameters of the class-conditional GMMs (2). For a predefined number of components M_c and parameterization of Σ_{cj} , the parameters μ_{cj} , Σ_{cj} and π_{cj} of the GMMs (2) are determined by an *expectation-maximization* (EM) algorithm. The EM algorithm is initialized by a clustering algorithm (we applied the k-means algorithm, described in Duda et al. 2001, chap. 10) which partitions the training data $\{\mathbf{x}_{c1}, \mathbf{x}_{c2}, \dots, \mathbf{x}_{cN_c}\}$, from each class $c=1, 2, \dots, C$ into M_c groups G_{cj} , $j=1, 2, \dots, M_c$. An indicator vector $\mathbf{z}_{cn} = \{z_{n1}^{(c)}, z_{n2}^{(c)}, \dots, z_{nM_c}^{(c)}\}$ is associated with each observation \mathbf{x}_{cn} . For $\mathbf{x}_{ci} \in G_{cj}$, $z_{nj}^{(c)} = 1$, otherwise $z_{nj}^{(c)} = 0$ for $n=1, \dots, N_c$ 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 a convergence criterion is satisfied (McLachlan 1997). Using the initial $\gamma_{nj}^{(c)} = z_{nj}^{(c)}$ and the initial $\lambda_{cj}=1$, for $j=1, \dots, M_c$, the following E- and M-steps are cycled for a predefined (fixed) number of components M_c and parameterization of Σ_{cj} .

M-step:

$$N_{cj} \leftarrow \sum_{n=1}^{N_c} \gamma_{nj}^{(c)}; \quad \mu_{cj} \leftarrow \frac{\sum_{n=1}^{N_c} \gamma_{nj}^{(c)} \mathbf{x}_{cn}}{N_{cj}}; \quad \pi_{cj} \leftarrow \frac{N_{cj}}{N_c};$$

$$\mathbf{W}_{cj} \leftarrow \frac{1}{N_{cj}} \sum_{n=1}^{N_c} \gamma_{nj}^{(c)} (\mathbf{x}_{cn} - \mu_{cj})(\mathbf{x}_{cn} - \mu_{cj})^T. \quad (3)$$

Calculate Σ_{cj} using the expressions (Celeux and Govaert 1995) listed in Table 1.

Table 1: Expressions for computation of Σ_{cj} .

Σ_{cj}	λ_{cj}	\mathbf{D}_c (or \mathbf{D}_{cj})	\mathbf{F}_c (or \mathbf{F}_{cj})
$\lambda_{cj}\mathbf{I}$	$tr(\mathbf{W}_{cj})/d$	-	-
$\lambda_{cj}\mathbf{D}_c$	$tr(\mathbf{W}_{cj}\mathbf{D}_{cj}^{-1})/d$	$\sum_j \lambda_{cj}^{-1} N_{cj} \text{diag}(\mathbf{W}_{cj})/N_c$	-
\mathbf{D}_{cj}	-	$\text{diag}(\mathbf{W}_{cj})$	-
$\lambda_{cj}\mathbf{F}_c$	$tr(\mathbf{W}_{cj}\mathbf{F}_c^{-1})/d$	-	$\sum_j \lambda_{cj}^{-1} N_{cj} \mathbf{W}_{cj}/N_c$
\mathbf{F}_{cj}	-	-	\mathbf{W}_{cj}

E-step:

$$\gamma_{nj}^{(c)} \leftarrow \frac{\pi_{cj} \Phi_{\Sigma_{cj}}(\mathbf{x}_{cn} - \boldsymbol{\mu}_{cj})}{\sum_{j=1}^{M_c} \pi_{cj} \Phi_{\Sigma_{cj}}(\mathbf{x}_{cn} - \boldsymbol{\mu}_{cj})} \quad (4)$$

2.2 BIC model selection

Following the paper of Fraley and Raftery (2002) we describe the model selection (setting the appropriate number M_c and parameterization Σ_{cj}) for GMM (2) based on the *Bayesian information criterion* (BIC) (Schwartz 1979), which has been found to work well in practical problems for discriminant analysis (Fraley and Raftery 1998, 2002; Halbe and Aladjem 2005). The BIC(M_c) for $f_c(\mathbf{x})$ (2) having M_c components is

$$\text{BIC}(M_c) = 2 \sum_{n=1}^{N_c} \ln f_c(\mathbf{x}_{cn}) - v_c \ln(N_c), \quad (5)$$

where $f_c(\mathbf{x}_{cn})$ denotes the GMM (2) fitted to the training data \mathbf{x}_{cn} , $n=1, 2, \dots, N_c$ ($\mathbf{x}_{cn} \in \mathbb{R}^d$) by the EM algorithm (for fixed M_c and Σ_{cj}) and v_c is the number of the adjustable parameters of the GMM (2). The large value of BIC(M_c) (5) corresponds to the model which is favored by the data. Table 2 gives the expressions for v_c , for the parameterizations of Σ_{cj} used in this paper.

Table 2: Number v_c of the adjustable parameters of d-variate GMM (2) having M_c Gaussian components with covariance matrix parameterization Σ_{cj} .

Σ_{cj}	v_c
$\lambda_{cj}\mathbf{I}$	$M_c+\theta$
$\lambda_{cj}\mathbf{D}_c$	$M_c+d+\theta$
\mathbf{D}_{cj}	$M_cd+\theta$
$\lambda_{cj}\mathbf{F}_c$	$M_c+d(d+1)/2+\theta$
\mathbf{F}_{cj}	$M_cd(d+1)/2+\theta$

In Table 2 $\theta=M_cd+M_c-1$ defines the summation of the number of components of the mean vectors μ_{cj} (M_cd) and the number of the independent mixing coefficients π_{cj} (M_c-1).

Using $\text{BIC}(M_c)$ (5) the number M_c and the parameterization of Σ_{cj} are set by a procedure proposed in our previous paper (Halbe and Aladjem 2005). First the appropriate number M_c of the mixture components is selected to correspond to the maximum of $\text{BIC}(M_c)$ (5) for each parameterization of Σ_{cj} . Then we select the parameterization of Σ_{cj} corresponding to the largest value of $\text{BIC}(M_c)$ (5).

Our procedure (Halbe and Aladjem 2005) is a modification of the original proposal of Fraley and Raftery (2002). It overcomes the problem in setting the maximal number $M_{c \max}$ of components required by Fraley and Raftery (2002) and reduces the computational complexity.

In the next section we propose a two step model selection procedure. In the first step we select the number of components M_c for $\Sigma_{cj}=\lambda_{cj}\mathbf{D}_c$ using the above-described procedure. In the second step we shrink $\lambda_{cj}\mathbf{D}_c$ to more complicated parameterizations of Σ_{cj} , which maximizes the cross-validation estimates of the correct classification rate.

3. Two-step model selection for the class-conditional GMMs

In our previous study (Halbe and Aladjem 2005) we observed that the simple (i.e., having a small number of adjustable parameters) parameterization $\Sigma_{cj}=\lambda_{cj}\mathbf{D}_c$ is an appropriate choice for wide classification applications. Moreover, the estimation of the parameters of $\lambda_{cj}\mathbf{D}_c$ is carried out by operations with diagonal matrices (see Table 1). The latter implies a stable EM algorithm for a small number of training observations and a large dimension of the observations.

In the EM algorithm (Section 2.1) the training observations, \mathbf{x}_{cn} , $n=1, 2, \dots, N_c$ ($\mathbf{x}_{cn}\in\mathbb{R}^d$), are partitioned into M_c clusters and $\boldsymbol{\mu}_{cj}$, Σ_{cj} for each cluster are estimated (Expression (3) and Table 1). For small N_c and large d the estimation of the unrestricted (full) Σ_{cj} , becomes highly variable or even unidentifiable (Friedman 1989). As a consequence the EM algorithm becomes unstable. This phenomenon becomes more pronounced as the sample size decreases.

Additionally, it was shown experimentally (Friedman 1989; Pima and Aladjem 2004) that the regularization method has the potential to increase the classification accuracy for a small training sample size.

Based on these observations we propose the following two step model selection procedure. In the first step we set the number M_c of the components of the GMM (2) for $\Sigma_{cj}=\lambda_{cj}\mathbf{D}_c$ by the model selection procedure based on BIC (Section 2.2). In the second step we shrink the predefined $\lambda_{cj}\mathbf{D}_c$ to a more complicated parameterization in the following way:

$$\Sigma_{cj}(\alpha_c) = (1 - \alpha_c)\lambda_{cj}\mathbf{D}_c + \alpha_c\mathbf{W}_{cj}, \quad \alpha_c \in [0,1]. \quad (6)$$

Here, \mathbf{W}_{cj} (3) is the scatter matrix computed in the last iteration of the EM algorithm (Section 2.1) for $\Sigma_{cj}=\lambda_{cj}\mathbf{D}_c$. The matrices \mathbf{W}_{cj} , $j=1, 2, \dots, M_c$, act as unrestricted (full)

component covariance matrices of the GMM (2). The parameter $\alpha_c \in [0, 1]$ controls the degree of shrinkage of the predefined covariance matrices $\lambda_{c_j} \mathbf{D}_c$ toward \mathbf{W}_{c_j} . The value $\alpha_c=1$ gives rise to \mathbf{W}_{c_j} , whereas $\alpha_c=0$ yields the predefined $\lambda_{c_j} \mathbf{D}_c$. Values between these limits compromise between the predefined diagonal matrices $\lambda_{c_j} \mathbf{D}_c$ and the unrestricted (full) covariance matrices \mathbf{W}_{c_j} .

Then we regularize the $\Sigma_{c_j}(\alpha_c)$ (6) through

$$\Sigma_{c_j}(\alpha_c, \beta_c) = (1 - \beta_c) \Sigma_{c_j}(\alpha_c) + \beta_c \sum_{j=1}^{M_c} \frac{N_{c_j}}{N_c} \Sigma_{c_j}(\alpha_c). \quad (7)$$

For a given value of α_c , the regularization parameter $\beta_c \in [0, 1]$ controls shrinkage toward the averaged matrix $\sum_{j=1}^{M_c} (N_{c_j} / N_c) \Sigma_{c_j}(\alpha_c)$ (often called the pooled scatter matrix). In our scenario for $\beta_c=1$, $\Sigma_{c_j}(\alpha_c, \beta_c)$ acts as common covariance matrix, which is the same for all the components of the GMM (2). This implies a considerable reduction of the degree of freedom of $\Sigma_{c_j}(\alpha_c)$ (6) by shrinkage toward the common matrix $\sum_{j=1}^{M_c} (N_{c_j} / N_c) \Sigma_{c_j}(\alpha_c)$. Even if the component covariance matrices $\Sigma_{c_j}(\alpha_c)$ are substantially different, the decrease in variability accomplished by using $\sum_{j=1}^{M_c} (N_{c_j} / N_c) \Sigma_{c_j}(\alpha_c)$ can sometimes lead to superior performance, especially in small-sample data setting (Section 5.5).

In summary, holding β_c fixed at 0 and varying α_c produces $\Sigma_{c_j}(\alpha_c)$ between the predefined diagonal $\Sigma_{c_j}=\lambda_{c_j} \mathbf{D}_c$ and the unrestricted (full) component covariance matrix \mathbf{W}_{c_j} . Holding α_c fixed and increasing β_c we shrink $\Sigma_{c_j}(\alpha_c)$ toward the common matrix $\sum_{j=1}^{M_c} (N_{c_j} / N_c) \Sigma_{c_j}(\alpha_c)$ (7). The matrix $\Sigma_{c_j}(\alpha_c, \beta_c)$, computed by (6) and (7), compromises between the diagonal parameterization $\lambda_{c_j} \mathbf{D}_c$ with a small number of $(d+1)$ adjustable parameters and the richer (unrestricted) parameterizations \mathbf{W}_{c_j} and

$\mathbf{W}_c = \sum_{j=1}^{M_c} (N_{cj} / N_c) \mathbf{W}_{cj}$ having $d(d+1)/2$ adjustable parameters. The expressions (6) and (7) imply a rich class of alternatives to the conventional GMM (2).

The second step of our method is stimulated by a method of Friedman, named *regularized discriminant analysis* (RDA) (Friedman 1989). In the next section we discuss the relation of our method to the RDA.

Ideally if we have enough training data, we could set aside a validation set and use it for setting the shrinkage parameters α_c, β_c suitable for class separation. Since training data are often scarce, this is usually not possible. In this case, we have to apply sample reuse methods - *cross-validation* (CV) methods (Stone 1974) or bootstrap methods (Efron 1983; Efron and Tibshirani 1993). The computational advantage associated with the CV stimulated us to apply it in our method.

In our implementation we use the *10-fold CV method* (Stone 1974; Hastie et al. 2001). We split the training set into ten roughly equal sized parts. Using nine parts we fit (by the EM algorithm, Section 2.1) the GMMs (2) for $\Sigma_{cj} = \lambda_{cj} \mathbf{D}_c$ and select the appropriate number M_c which corresponds to the maximum of $\text{BIC}(M_c)$ (5) (Section 2.2). Then we calculate the correct classification rate, allocating the observations from the left part (acting as a validation set) by the plug-in Bayes classification rule (Section 1). We iterate the computation of the correct classification rate at a grid of points on the α_c, β_c plane ($\alpha_c, \beta_c \in [0, 1]$) using (5) and (7). We replicate this process ten times using different parts for training and validation each time. Then we average the correct classification rates over the 10 replications for each α_c, β_c and choose the point $\hat{\alpha}_c, \hat{\beta}_c$ with the maximum averaged rate. The point $\hat{\alpha}_c, \hat{\beta}_c$ defines the final solution of our

method – the class-conditional GMM (2) having component covariance matrices $\Sigma_{c_j}(\hat{\alpha}_c, \hat{\beta}_c)$.

It is possible that several values of $\hat{\alpha}_c, \hat{\beta}_c$ provide the same value of the CV correct classification rate. In this case we select the point $(\hat{\alpha}_c, \hat{\beta}_c)$ which corresponds to the most parsimonious model, i.e. the point $(\hat{\alpha}_c, \hat{\beta}_c)$ which favors the parameterizations $\lambda_c \mathbf{D}_c$ and \mathbf{W}_c having reduced degrees of freedom.

In our experiments we used an optimization grid defined to have 50 points by the outer product of $\alpha_c=(0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.95)$ and $\beta_c=(0, 0.125, 0.354, 0.650, 1)$. The latter implies 50^C possible parameterizations $\Sigma_{c_j}(\alpha_c, \beta_c)$ for a C-class classification problem. Computationally more practical is a procedure resulting in equal values for $\hat{\alpha}, \hat{\beta}$ for the classes. In this procedure we search over 50 parameterizations instead over 50^C parameterizations for the exhaustive search. In Section 5.1 we studied this procedure for four data sets having $C=2$ and obtained a negligible difference in performance compared to the exhaustive search. This result motivated the application of the latter search procedure in the rest of experiments presented in Section 5.

4. Relation to Friedman’s regularized discriminant analysis

As we mentioned in previous section our method is stimulated by and closely related to Friedman’s (1989) *regularized discriminant analysis* (RDA). Here we explain the similarities and differences of the methods.

Friedman (1989) proposed a regularization method to mitigate the problem of highly variable or even unidentifiable covariance matrices for quadratic discriminant analysis. He studied the plug-in-Bayes classification rule for $f_c(\mathbf{x}) = \Phi_{\Sigma_c}(\mathbf{x} - \boldsymbol{\mu}_c)$ having

unrestricted Σ_c for the classes $c=1, 2, \dots, C$. The RDA shrinks the unrestricted Σ_c toward a common covariance

$$\Sigma = \sum_{c=1}^C \frac{N_c}{N} \Sigma_c, \quad (8)$$

where $N = \sum_{c=1}^C N_c$ is the number of the training observations for all classes. The regularized matrix is obtained by the following expression:

$$\Sigma_c(\alpha) = \frac{(1-\alpha)N_c}{(1-\alpha)N_c + \alpha N} \Sigma_c + \frac{\alpha N}{(1-\alpha)N_c + \alpha N} \Sigma, \quad \alpha \in [0, 1], \quad (9)$$

Comparing (9) to our proposal (6) we observe that the way of shrinkage of the covariance matrices is similar. The main difference is in the meaning of the shrinkage. Our method operates on the matrices computed within class c (see expression (6)), while Friedman's method operates on the matrices from different classes (Σ (8) is the average of the class-conditional covariance matrices Σ_c , $c=1, 2, \dots, C$). In the RDA the migration (9) is from $\Sigma_1 \neq \Sigma_2 \neq \dots \Sigma_C$ (quadratic discriminant analysis) to $\Sigma_c = \Sigma$, $c=1, 2, \dots, C$, (linear discriminant analysis). In our method the migration (6) is from $\lambda_{c_j} \mathbf{D}_c$ (GMM with equal diagonal covariance matrix \mathbf{D}_c for all components) to different \mathbf{W}_{c_j} (GMM with different unrestricted component matrices). In addition, instead of Friedman's weightings $(1-\alpha)N_c / [(1-\alpha)N_c + \alpha N]$ and $\alpha N / [(1-\alpha)N_c + \alpha N]$ in (9), we apply just the weightings $(1-\alpha)$ and α in (6). By this we overcome the setting of a non-equally spaced grid for the values of α needed in the method of Friedman (1989, eq. 16a), which is caused by the nonlinear behavior of Friedman's weightings, illustrated in Figure 1 for $N_c=50$ and $N=350$.

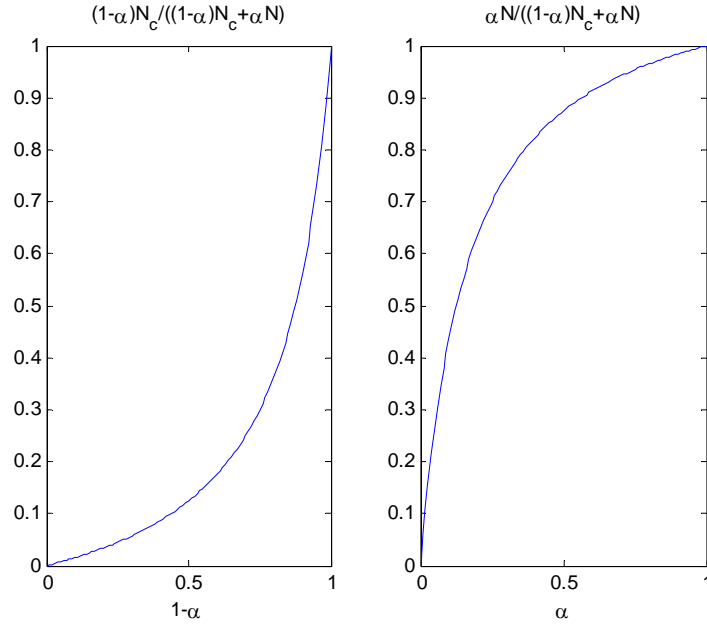


Figure 1: Non-linear behavior of Friedman's weights ($N_c=50$ and $N=350$).

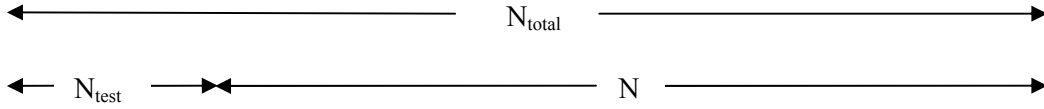
Our regularization expression (7) is closely related to the proposal of Friedman (1989, eq. 18). The goal in the both methods is to approach the parameterization of the covariance matrix with substantially reduced degrees of freedom (reduced number of the adjustable parameters). As previously, the difference is in the meaning of the shrunk matrices. Our method operates on the matrices for a certain class c while RDA shrinks matrices from different classes.

5. Experiments

In this section we study the correct classification rate of the plug-in Bayes classifier (Section 1) for class-conditional GMMs (2) having M_c Gaussian components and parameterization of Σ_{cj} set by different *model selection* (MS) methods. We compare our method (Section 3) with the method of Fraley and Raftery (2002), named MS_BIC. The method MS_BIC selects M_c and the parameterization of Σ_{cj} which correspond to the maximal value of $BIC(M_c)$ (5). We have explained MS_BIC in Section 2.2. In addition

we study a MS method, named MS_BIC_CV which selects M_c corresponding to the maximal value of $BIC(M_c)$ (5) (as in the method MS_BIC) but the parameterization of the Σ_{cj} is selected by 10-fold CV, as we select $\hat{\alpha}$, $\hat{\beta}$ in our method (see the end of Section 3). Finally, we select the GMM by a pure CV method, named MS_CV_CV. The method MS_CV_CV selects both M_c and Σ_{cj} by 10-fold CV.

We carried out experiments on various artificial and real-world data sets, explained in Section 5.1. For each data set we replicated the following procedure 10 times. We drew observations with replacement (from the original data) and rotated the data set randomly. Then we split the available data set (having N_{total} number of observations) into two subsets – training set and test set (see Figure 3). The training set is used for fitting the class-conditional GMMs (2) (by the EM algorithm, Section 2.1) and for setting the appropriate M_c and Σ_{cj} by the compared methods MS_BIC, MS_BIC_CV, MS_CV_CV and our method (Section 3). We separated validation set from the training set for setting the appropriate M_c and Σ_{cj} by MS_CV_CV. We set the maximum number $M_{c\ max}=15$ of the Gaussian components. For MS_BIC_CV and our method we used the validation set for setting Σ_{cj} . We applied the 10-fold CV, illustrated in Figure 2. The test set was used for computation of the classification performance of MS_BIC, MS_BIC_CV, MS_CV_CV and our method. We evaluated the classification performance by the percentage of test observations allocated by the plug-in Bayes classification rule correctly. We will refer to this percentage as test correct classification rate. We computed this rate by 10-fold CV (illustrated in Figure 2). Finally, we averaged the test correct classification rate over 10 random drawings of the data sets from the original data.



1	Test set	Training set for EM algorithm and BIC model selection (MS_BIC)	
		Validation set for model selection (MS_BIC_CV, MS_CV_CV and our method)	Training set for EM algorithm (MS_BIC_CV, MS_CV_CV and our method)
2	Test set	Training set for EM algorithm and BIC model selection (MS_BIC)	
		Validation set for model selection (MS_BIC_CV, MS_CV_CV and our method)	Training set for EM algorithm (MS_BIC_CV, MS_CV_CV and our method)
10	Test set	Training set for EM algorithm and BIC model selection (MS_BIC)	
		Training set for EM algorithm (MS_BIC_CV, MS_CV_CV and our method)	Validation set for model selection (MS_BIC_CV, MS_CV_CV and our method)
11	Test set	Training set for EM algorithm and BIC (MS_BIC)	
		Validation set for model selection (MS_BIC_CV, MS_CV_CV and our method)	Training set for EM algorithm (MS_BIC_CV, MS_CV_CV and our method)
100	Test set	Training set for EM algorithm and BIC model selection (MS_BIC)	
		Validation set for model selection (MS_BIC_CV, MS_CV_CV and our method)	Training set for EM algorithm (MS_BIC_CV, MS_CV_CV and our method)

Figure 2: Illustration of the partition of a data set for the 10-fold CV.

As we mentioned in Section 1, the pure CV selection by MS_CV_CV is not practical for large number of the classes. In our study for five parameterizations of Σ_{c_j} and $M_{c_{\max}}=15$ we have to consider $(5 \times 15)^C$ models for each CV computation. As a result we have to consider $(5 \times 15)^2=5625$ models for two class problem ($C=2$); $(5 \times 15)^3=421875$ models for $C=3$; 31×10^6 models for $C=4$ and 23×10^8 models for $C=5$. In order to keep a reasonable complexity of the experiments we chose to run MS_CV_CV for small sample size data sets and $C=2$ (Section 5.5).

5.1 Data sets

We carried out experiments on various artificial and real-world data sets (listed in Table 3), out of the UCI Machine Learning Repository (Blake, Keogh and Merz 1998) and G. Ratsch benchmark data sets (<http://ida.first.fraunhofer.de/~raetsch>). In addition to those data sets we composed a data set named MODIFIED LETTER. It merges the classes (26 letters) of the benchmark data set LETTER (Blake et al. 1998) into two groups. We set letters O, U, P, S, X, Z, E, B, F, T, W, A, Q to be group one and the letters H, D, N, C, R, G, K, Y, V, M, I, J, L to be group two, using an agglomerative hierarchical clustering (Duda et al. 2001) with between-groups linkage of the means of the letters. The goal was to set complicated (highly overlapped) groups. In Figure 3 we show the dendrogram for the means of the letters, where the bold letters are the letters of group one. The first level of the dendrogram shows the 26 means of the letters as singleton clusters. At the second level, the close means of the letters have been grouped to form a cluster, and they stay at all subsequent levels. The dendrogram shows the similarity between the clusters that are grouped. As can be seen from the dendrogram, we set highly overlapped groups, having letters from different clusters. The MODIFIED

LETTER data set contains 375 observations drawn from the original LETTER data set randomly.

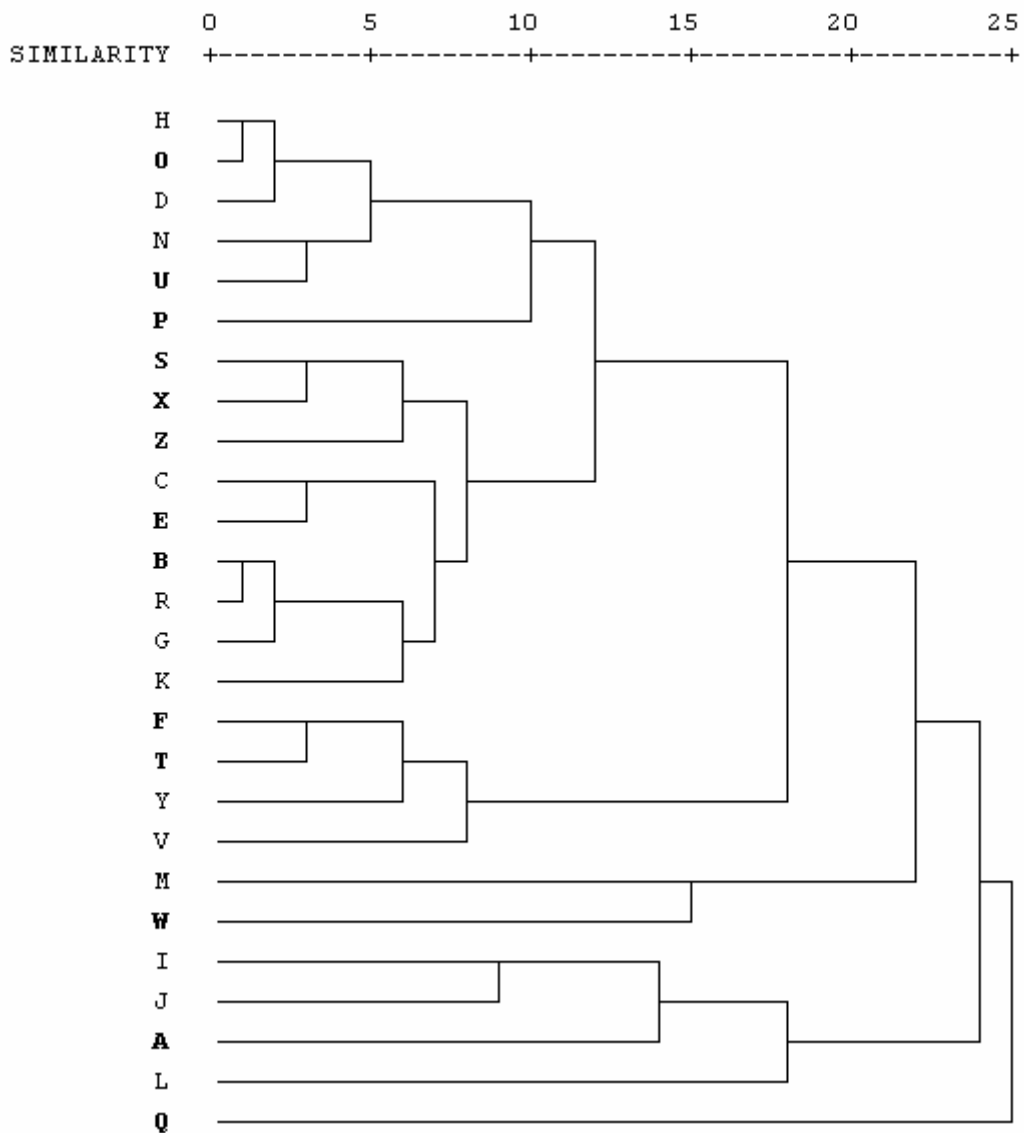


Figure 3 : Dendrogram from agglomerative hierarchical clustering of means of the LETTER data set.

Table 3 shows the characteristics of the data sets used in the experimental study. N_{total} denotes the available number of the observations; N is the number of the training observations; N_{test} is the number of the test observations; C denotes the number of the classes; p is the original dimension of the observations and d is the (used in the

experiments) actual dimension of the observations obtained by principal component analysis (PCA) dimension reduction (Duda et al. 2001; Jolliffe 2002). We retain with 5% error of the reduction by PCA. S denotes the averaged ratio of the number of the observations in each class to the dimension of the observations, i.e.:

$$S = \sum_{c=1}^C \hat{P}_c \frac{N_c}{d}. \quad (10)$$

Here N_c is the number of the training observations for class c , d is the dimension of the observations and $\hat{P}_c = N_c / \sum_{c=1}^C N_c$.

Table 3: Data set characteristics.

Data	N_{total}	N	N_{test}	C	p	d	S
SONAR	208	188	20	2	60	29	3.26
WINE	178	161	17	3	13	12	4.58
IONOSPHERE	350	315	35	2	34	30	5.66
GLASS	214	193	21	6	9	6	8.47
VOWEL	990	891	99	11	13	9	9
MODIFIED LETTER	375	338	19662	2	16	15	11.51
IRIS	150	135	15	3	4	3	15
WAVEFORM NOISE	5000	4500	500	3	40	40	37.50
LETTER	$2 \cdot 10^4$	18000	2000	26	16	15	46.20
WAVEFORM	5000	4500	500	3	21	21	71.44
BANANA	5002	4502	500	2	2	2	1138
BANANA NOISE	5002	4502	500	2	2	2	1138

As can be seen from Table 3, the data sets differ largely in dimension and number of observations and cover a wide spectrum of data sets.

5.2 **Comparison of the exhaustive search and the search resulting equal values α , β for the classes**

As we explained in the end of Section 3 in order to restrict the computation to a reasonable complexity we carried out most of the experimental studies of our method for equal values of the shrinkage parameters α and β for the classes. In this section we

justify the reasonableness of this procedure by a comparison with the exhaustive search for $(\hat{\alpha}_c, \hat{\beta}_c)$. We carried out experiments with data sets having two classes ($C=2$).

In Table 4 we report the obtained results. The first column contains the list of the data sets used in the experiment. The second column gives the results for the search resulting in equal values for α , β for the classes and the third column gives the results for the exhaustive search. For each data set we report the test correct classification rate together with the standard error. In parenthesis we report the averaged values of the shrinkage parameters. The last column gives the significance difference level (Sig.) between the classification rates for the search resulting in equal values of α and β for the classes (column 2) and for the exhaustive search (column 3). We computed the significance of the difference by a paired t-test (Robert and Torrie 1980). We observed that for all data sets this difference is insignificant (Sig.>0.05). Consequently we obtained negligible difference of the classification rates. Based on this result we decided to carry out the rest of the experiments using the procedure resulting in equal values of α and β for the classes, which reduced the computational complexity significantly.

Table 4: Test correct classification rates for our method for different search procedures for the shrinkage parameters.

DATA	Search resulting in equal $(\hat{\alpha}, \hat{\beta})$ for the classes	exhaustive search for $(\hat{\alpha}_1, \hat{\beta}_1), (\hat{\alpha}_2, \hat{\beta}_2)$,	Sig.
SONAR	80.05±0.78 (0.65,0.34)	79.70±0.87 (0.53,0.38); (0.67,0.41)	0.597
IONO-SPHERE	93.63±0.44 (0.16,0.22)	93.66±0.44 (0.46,0.29); (0.17,0.25)	0.798
BANANA	94.93±0.08 (0.86,0.08)	94.92±0.08 (0.79,0.21); (0.89,0.03)	0.865
BANANA-NOISE	75.26±0.16 (0.73,0.12)	75.34±0.16 (0.71,0.28); (0.56,0.07)	0.091

5.3 Comparison of our method with conventional GMMs

Here we compare our method (Section 3) with conventional GMMs having the parameterizations $\Sigma_{cj}=\lambda_{cj}\mathbf{I}$, $\lambda_{cj}\mathbf{D}_c$, \mathbf{D}_{cj} , $\lambda_{cj}\mathbf{F}_c$ and \mathbf{F}_{cj} . For the conventional GMMs we set the number M_c of the components corresponding to a maximal value of $\text{BIC}(M_c)$ (5) for each Σ_{cj} . In this section we don't set Σ_{cj} by a model selection method but just compare our method to the conventional GMM having the largest test classification rate. Consequently, in this section we favor the conventional GMMs using their largest test rates in the comparison. In Table 5 we report the results obtained.

Table 5: Test correct classification rates for the regularized GMM (our method) and conventional GMMs.

DATA	Regularized GMM ($\hat{\alpha}, \hat{\beta}$) values	Conventional GMMs having different parameterizations of the component covariance matrices Σ_{cj}				
		$\lambda_{cj}\mathbf{I}$	$\lambda_{cj}\mathbf{D}_c$	\mathbf{D}_{cj}	$\lambda_{cj}\mathbf{F}_c$	\mathbf{F}_{cj}
SONAR	80.05±0.78 (0.65, 0.34)	71.00±0.88	73.20±0.83	72.80±0.91	79.05±0.91	80.00±0.84 Sig=0.802
WINE	98.71±0.26 (0.56, 0.65)	92.53±0.73	93.00±0.63	92.88±0.59	98.94±0.22 Sig=0.184	98.88±0.25
IONO-SPHERE	93.60±0.45 (0.18, 0.23)	92.09±0.44	92.89±0.46 Sig=0.015*	92.26±0.48	92.17±0.46	84.40±0.64
GLASS	68.62±0.97 (0.59, 0.03)	65.40±1.21	65.19±1.15	66.45±1.11	67.30±0.99 Sig=0.254	67.30±0.99 Sig=0.254
VOWEL	97.71±0.40 (0.70, 0.14)	86.09±0.44	87.24±0.45	88.44±0.45	89.73±0.44 Sig=0.000	89.73±0.44 Sig=0.000*
MODIFIED LETTER	82.47±0.13 (0.78, 0.30)	75.61±0.18	75.64±0.16	75.38±0.16	80.32±0.13 Sig=0.000*	79.61±0.11
IRIS	97.00±0.43 (0.46, 0.52)	96.53±0.49	96.93±0.46	96.20±0.53	97.47±0.35	97.60±0.35 Sig=0.012*
WAVEFORM-NOISE	86.63±0.15 (0.07, 0.15)	86.64±0.15	86.69±0.15 Sig=0.451	86.62±0.15	84.22±0.16	84.22±0.16
LETTER	96.67±0.05 (0.95, 0.13)	80.41±0.19	82.45±0.20	83.77±0.21	91.30±0.10	94.63±0.09 Sig=0.000*
WAVEFORM	86.28±0.12 (0.07, 0.16)	86.60±0.12 Sig=0.000*	86.35±0.13	86.46±0.13	85.31±0.15	85.06±0.15
BANANA	94.92±0.08 (0.86, 0.08)	92.94±0.14	94.24±0.10	94.70±0.10	94.40±0.11	95.06±0.09 Sig=0.014*
BANANA-NOISE	75.25±0.16 (0.72, 0.11)	75.34±0.17	75.20±0.17	75.33±0.17	75.62±0.18	75.82±0.17 Sig=0.000*

As previously, we list the test correct classification rates along with standard errors. The largest test rates for the conventional GMM are written in bold. Under the largest rates we report the significance level (Sig.) of the difference between these rates and the test rates of the regularized GMMs (our method from Section 3). As previously, we used the paired t-test for computation of the significance level (Sig.). We indicate by star (*) the significant differences (Sig<0.05).

Our method implies significant enlargement (with respect conventional GMMs) of the test rates for data sets (written in bold in Table 5) IONOSPHERE (enlargement 0.71%), VOWEL (enlargement 7.98%), MODIFIED LETTER (enlargement 2.15%) and LETTER (enlargement 2.04%). This shows that our method has the potential to produce parameterizations $\Sigma_{cj}(\hat{\alpha}, \hat{\beta})$ which are better than the five basic parameterizations $\Sigma_{cj}=\lambda_{cj}\mathbf{I}$, $\lambda_{cj}\mathbf{D}_c$, \mathbf{D}_{cj} , $\lambda_{cj}\mathbf{F}_c$ and \mathbf{F}_{cj} . It is not surprising that the conventional GMM is better for some data sets, i.e., WAVEFORM (decrease 0.32%), BANANA (decrease 0.14%) and BANANA-NOISE (decrease 0.52%) (notice the small decrease in the classification rates of our method for WAVEFORM, BANANA and BANANA-NOISE). As we mentioned we favored the conventional GMM by selecting the best parameterization (written in bold) using the test data, while the regularized GMM (our method) set the $\hat{\alpha}, \hat{\beta}$ (the parameterization $\Sigma_{cj}(\hat{\alpha}, \hat{\beta})$) using training data only.

Comparing the test rates of the regularized GMM (having predefined parameterization $\lambda_{cj}\mathbf{D}_c$, Section 3) and the conventional GMM with $\Sigma_{cj}=\lambda_{cj}\mathbf{D}_c$ we observe significant improvement of the test rate for the regularized GMM. This indicates the benefit of using our method.

Studying the values for $\hat{\alpha}$ in Table 5 we observe the natural behavior of the shrinkage. For the data sets which favor the complex parameterizations $\lambda_{cj}\mathbf{F}_c$ and \mathbf{F}_{cj}

(SONAR, WINE, GLASS, VOWEL, MODIFIED LETTER, LETTER, BANANA-NOISE) the values of $\hat{\alpha}$ tend to be large, favoring the complex parameterizations \mathbf{W}_{c_j} in (6). For the data sets which favor simple models $\lambda_{c_j}\mathbf{I}$, $\lambda_{c_j}\mathbf{D}_c$ and \mathbf{D}_{c_j} (IONOSPHERE, WAVEFORM-NOISE, WAVEFORM) we observe small value of $\hat{\alpha}$ which favor the simple component $\lambda_{c_j}\mathbf{D}_c$ in (6). The behavior of the values of $\hat{\beta}$ is similar.

5.4 Comparison of our method with the methods MS_BIC and MS_BIC_CV

In the previous section we compared the classification accuracy of our method (Section 3) and the conventional GMMs with parameterizations $\Sigma_{c_j}=\lambda_{c_j}\mathbf{I}$, $\lambda_{c_j}\mathbf{D}_c$, \mathbf{D}_{c_j} , $\lambda_{c_j}\mathbf{F}_c$ and \mathbf{F}_{c_j} favoring the conventional GMMs (using the test data for setting Σ_{c_j}). Here we study the conventional GMM in a more realistic (for the applications) situation. We select the parameterization of Σ_{c_j} by the methods MS_BIC or MS_BIC_CV (using the training data).

In Table 6 we report the results obtained. The column for regularized GMM (our method) is copied from Table 5 and the last columns contain the results for the methods MS_BIC_CV and MS_BIC. As previously, we write the largest test rates for MS_BIC_CV and MS_BIC in bold and report the significance level (Sig.) of the difference between these rates and the test rates of the regularized GMM. We indicated by star (*) the significant difference (Sig<0.05). Finally, S is the ratio of the number of training observations to the data dimension, which is copied from Table 3.

Table 6: Test correct classification rates for regularized GMM (our method), MS_BIC_CV and MS_BIC.

DATA	S	Regularized GMM (α, β) values	GMMs for MS_BIC_CV	GMMs for MS_BIC
SONAR	3.26	80.05±0.78 (0.65, 0.34)	79.45±0.88 Sig=0.359	79.05±0.90
WINE	4.58	98.71±0.26 (0.56, 0.65)	99.05±0.23 Sig=0.096	97.76±0.35
IONOSPHERE	5.66	93.60±0.45 (0.18, 0.23)	92.17±0.42	92.34±0.50 Sig=0.003*
GLASS	8.47	68.62±0.97 (0.59, 0.03)	65.03±1.12	66.71±1.01 Sig=0.096
VOWEL	9	97.71±0.40 (0.70, 0.14)	89.75±0.40 Sig=0.000*	87.26±0.48
MODIFIED LETTER	11.51	82.47±0.13 (0.78, 0.30)	79.88±0.18	80.25±0.14 Sig=0.000*
IRIS	15	97.00±0.43 (0.46, 0.52)	96.87±0.48	97.27±0.40 Sig=0.207
WAVEFORM-NOISE	37.50	86.63±0.15 (0.07, 0.15)	86.59±0.15	86.64±0.15 Sig=0.981
LETTER	46.20	96.67±0.05 (0.95, 0.13)	94.63±0.09 Sig=0.000*	94.17±0.09
WAVEFORM	71.44	86.28±0.12 (0.07, 0.16)	86.48±0.13	86.60±0.12 Sig=0.000*
BANANA	1138	94.92±0.08 (0.86, 0.08)	95.05±0.09	95.06±0.09 Sig=0.014*
BANANA-NOISE	1138	75.25±0.16 (0.72, 0.11)	75.78±0.17	75.82±0.17 Sig=0.000*

We observed significant enlargement of test rates for our method with respect to MS_BIC and MS_BIC_CV for data sets (written in bold in Table 6) IONOSPHERE (enlargement 1.26%), VOWEL (enlargement 7.96%), LETTER (enlargement 2.04%) and MODIFIED LETTER (enlarging 2.22%). This result is not surprising. The goal of the regularization is to overcome overfitting, which is more pronounced as the sample size decreases and our method implies a significant improvement for data sets having small ratio S of the number of training observations to the data dimension. In the next

section we study this phenomenon more thoroughly. There is a small loss in applying our method for large sample sizes (WAVEFORM (decrease 0.32%), BANANA (decrease 0.14%), BANANA-NOISE (decrease 0.57%)).

Comparing the results for MS_BIC (Table 6) with the best test results of the conventional GMMs (Table 5) we observe that MS_BIC performs well for the data sets WAVEFORM, BANANA and BANANA-NOISE having a large sample size ($S > 10$). Consequently, the model selection based on BIC could be advised for practical application in the large sample size classification problems. The latter observation is consistent with the result in the paper of Biernacki and Govaret (1999). The conventional GMMs having an unrestricted (full) covariance matrix (see Table 5) outperform MS_BIC results for data sets SONAR, WINE, IONOSPHERE, GLASS, VOWEL, MODIFIED LETTER, IRIS and LETTER. The latter implies that BIC underestimates the complexity of $f_c(\mathbf{x})$ i.e., BIC penalizes complex models too heavily, giving preference to simpler models. Additionally, we observe that the results for MS_BIC_CV are quite similar to the MS_BIC results, but the cross-validation approach involves a greater computational load (see Section 3).

5.5 Small sample size studies

In this section we replicated the experiments explained in Section 5.4 for reduced values for the ratio S (10) of the number N of training observations to the data dimension for all data sets. The new values for S are listed in the last column of Table 7. Comparing the reduced ratios S (Table 7) with the original ratios S (Table 3) we observe a significant reduction of the new setting for the ratio S . In Table 7 we give the number of observations (N_{total} , N and N_{test}) for the new setting for S . In this experiment

the dimensionality d of the observations and the number C of the classes for the data sets were kept as previously (Table 3).

Table 7: Small sample size data sets characteristics.

DATA	N_{total}	N	N_{test}	S
SONAR	208	130	78	2.25
WINE	178	119	59	3.39
IONOSPHERE	350	189	161	3.40
GLASS	214	131	83	5.75
VOWEL	990	558	432	5.64
MODIFIED LETTER	375	133	19867	4.53
IRIS	150	51	99	5.67
WAVEFORM-NOISE	5000	360	4640	3
LETTER	$2 \cdot 10^4$	2338	17662	6
WAVEFORM	5000	355	4645	5.64
BANANA	5002	27	4975	6.82
BANANA-NOISE	5002	27	4975	6.82

Here, for data sets having two classes ($C=2$) we run `MS_CV_CV` in addition to other methods (our regularized GMM method, Section 3; `MS_BIC`, Fraley and Raftery 2002 and `MS_BIC_CV`). In Table 8 we report the obtained results.

Comparing the results of Table 8 and Table 6 we observe that the benefit of using our method is greater for reduced S . For the original size of data sets (Table 6) our method significantly outperforms the methods `MS_BIC` and `MS_BIC_CV` for data sets - IONOSPHERE (enlargement 1.26%), VOWEL (enlargement 7.96%), MODIFIED LETTER (enlargement 2.22%) and LETTER (enlargement 2.04%). While reducing S (Table 8) our method performs significantly better for more data sets – SONAR (enlargement 1.11%), IONOSPHERE (enlargement 0.64%), GLASS (enlargement 1.72%), VOWEL (enlargement 5.39%), MODIFIED LETTER (enlargement 1.03%) and LETTER (enlargement 4.73%). Additionally, the standard errors of our method are smaller (for almost all data sets) than those of the other methods. The latter implies that our method is less sensitive to the training data variability. Comparing `MS_BIC_CV`

with MS_BIC results we conclude that MS_BIC_CV (CV selection of Σ_{cj}) outperform MS_BIC (BIC selection of Σ_{cj}) for small sample size data sets. In some cases (data sets BANANA and IONOSPHERE) the method MS_CV_CV (pure CV model selection) produces results comparable to these of our method.

Table 8: Test correct classification rates for regularized GMM (our method), MS_BIC_CV, MS_BIC and MS_CV_CV for reduced ratio S.

DATA	S	Regularized GMMs	GMMs using MS_BIC_CV	GMMs using MS_BIC	GMMs using MS_CV_CV
SONAR	2.25	77.92±0.42 (0.58, 0.33)	73.91± 0.61	76.81±0.45 Sig=0.027*	74.38±0.61
WINE	3.39	97.07±0.29 (0.72, 0.54)	97.20±0.29 Sig=0.576	95.49±0.33	-
IONOSPHERE	3.40	93.61±0.16 (0.27, 0.23)	92.19±0.25	88.14±0.74	92.97±0.20 Sig=0.033*
GLASS	5.75	64.87±0.49 (0.61, 0.01)	63.15±0.60 Sig=0.034*	59.23±0.62	-
VOWEL	5.64	93.30±0.17 (0.71, 0.22)	87.91±0.23 Sig=0.000*	87.40±0.25	-
MODIFIED LETTER	4.53	76.75±0.21 (0.58, 0.35)	74.70±0.28	75.72±0.24 Sig=0.000*	75.22±0.32
IRIS	5.67	95.74±0.17 (0.75, 0.38)	95.37±0.16	95.60±0.22 Sig=0.543	-
WAVEFORM-NOISE	3	83.46±0.08 (0.07, 0.06)	83.75±0.07	84.14± 0.06 Sig=0.000*	-
LETTER	6	87.89± 0.06 (0.92, 0.34)	83.16±0.06 Sig=0.000*	82.71±0.09	-
WAVEFORM	5.64	82.16±0.09 (0.13, 0.18)	83.32±0.11	83.71±0.09 Sig=0.000*	-
BANANA	6.82	74.89±0.69 (0.80, 0.32)	72.87±0.66	73.32±0.70	75.61±0.63 Sig=0.011*
BANANA-NOISE	6.82	63.16±0.50 (0.79, 0.28)	62.86±0.54 Sig=0.365	61.90±0.53	-

6 Conclusions

In this paper we proposed a method for *mixture discriminant analysis* (MDA), assuming a *Gaussian mixture model* (GMM) for class-conditional density functions. We proposed (Section 3) a procedure for determining the parameterization of the covariance matrices of the GMM components. It compromises between the simplicity of the procedure of Fraley and Raftery (2002) based on the *Bayesian information criterion* (BIC) and the high accuracy of *cross-validation* (CV) procedure (Stone 1974; Hastie et al. 2001).

Our method is stimulated and closely related to the *regularized discriminant analysis* (RDA) proposed by Friedman (1989). We apply an idea of Friedman to shrink a predefined covariance matrix to the parameterizations with substantially reduced degrees of freedom (reduced number of the adjustable parameters). Our method differs from original RDA in the meaning of the shrinkage. We operate on the matrices computed for a certain class while RDA shrinks matrices from different classes.

We carried out an extensive experimental study (Section 5) of our method. We compare the correct classification rate of the plug-in Bayes classification using class-conditional GMMs set by our method (Section 3) with the classification rate for the conventional GMMs having fixed parameterizations of the component covariance matrices and with the classification rate for the GMMs set by the method MS_BIC (Fraley and Raftery 2002) and two implementations MS_BIC_CV and MS_CV_CV of CV model selection (Hastie et al. 2001). In the comparative study we used various artificial and real-world data sets which differ largely in the dimension and the number of observations and cover a wide spectrum of data sets. The results obtained show that our method has the potential to produce parameterizations of the component covariance

matrices which are better than the parameterizations usually used in the conventional GMMs. We observed significant enlargement of the correct classification rates for our method compared to other methods (MS_BIC_CV, MS_BIC and MS_CV_CV) for most of the data sets. This was more pronounced as the data sample size decreased. The latter shows that our method could be an attractive choice for applications based on a small number of training observations.

Acknowledgements

The authors wish to thank associated editor Prof. Robert P. W. Duin and the reviewers for their critical reading of the manuscript and helpful comments. This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University of the Negev, Israel.

References

- Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press.
- Banfield, J. D., and Raftery, A. E. (1993), "Model-based Gaussian and non Gaussian clustering," *Biometrics*, 49, 803-821.
- Bensmail, H., and Celeux, G. (1996), "Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition," *Journal of the American Statistical Association*, 91, 1743-1748.
- Biernacki, C., and Govaret, G. (1999), "Choosing Models in Model-based Clustering and Discriminant Analysis," *Journal of the Statistical Computation and Simulation*, 64, 49-71.

- Biernacki, C., Celeux, G, and Govaret, G. (2000), "Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719-725.
- Blake C., Keogh, E., and Merz, C. J. (1998), "UCI repository of machine learning databases." Irvin: University of California, Department of Information and Computer Sciences. Available in: <http://www.ics.uci.edu/~mlearn/>.
- Celeux, G., and Govaert, G. (1995), "Gaussian parsimonious clustering models," *Pattern Recognition*, 28, 781-793.
- Duda, R.O., Hart, P.E., and Stork, D.G. (2001), *Pattern Classification* (2nd ed.), John Wiley and Sons.
- Dempster, A. P., Laird, N., and Rubin, D. (1977), "Maximum Likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39, 1-38.
- Efron, B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American Statistical Association*, 78, 316-331.
- Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.
- Friedman, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165-175.
- Fraley, C., and Raftery, A. E. (1998), "How Many clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578-588.
- (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611-631.

- Halbe, Z., and Aladjem, M. (2005), "Model-based Mixture Discriminant Analysis-An Experimental Study," *Pattern Recognition*, 38, 437-440.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.
- Jolliffe, I. T. (2002), *Principal Component Analysis* (2th ed.), Springer-Verlag, New York.
- Kass, R. E., and Raftery, A. E (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773-795.
- McLachlan, G. J. (2004), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley.
- (1997), *The EM Algorithm and Extensions*, Wiley.
- McLachlan, G. J., and Basford, K. E. (1998), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.
- McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.
- Nabney, I. T. (2002), *NETLAB, Algorithms for Pattern Recognition*, Springer.
- Pima, I., and Aladjem, M. (2004), "Regularized discriminant analysis for face recognition," *Pattern Recognition* , 37, 1945-1948.
- Robert, R. G. S., and Torrie, J. H. (1980), *Principles and procedures of statistics a biomedical approach*, McGraw-Hill Book Company (2nd ed.).
- Schwartz, G. (1979), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.
- Stone, M. (1974), "Cross-Validation Choice and assessment of statistical prediction," *Journal of the Royal Statistical Society*, 36, 111-147.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Wiley & Sons, New York.