Semiparametric discriminant analysis for mixture populations

Subhajit Dutta[a]*and Probal Chaudhuri[a]

Mahalanobis distance and Fisher's linear discriminant analysis are fundamentally related to each other, and both the ideas have been extensively used in the classification literature. Fisher's linear and quadratic discriminant functions are known to possess Bayes risk optimality when the class distributions are Gaussian. However, they may have poor performance when the assumption of Gaussianity is violated, and the class boundaries for different populations have complex geometric shapes. In such situations, many standard nonparametric classifiers also have high misclassification rates when the data dimension is large due to the curse of dimensionality that affects such nonparametric methods. In this article, we derive an expression for the Bayes classifier when the class distributions are finite mixtures of elliptic distributions. The Bayes classifier turns out to be a function of the Mahalanobis distances of a data point from different sub-populations within a mixture population. We use this to develop and investigate a semi-parametric classification procedure, which is named as SPARC (SemiPARametric Classification), and we demonstrate its performance on simulated and real benchmark data sets.

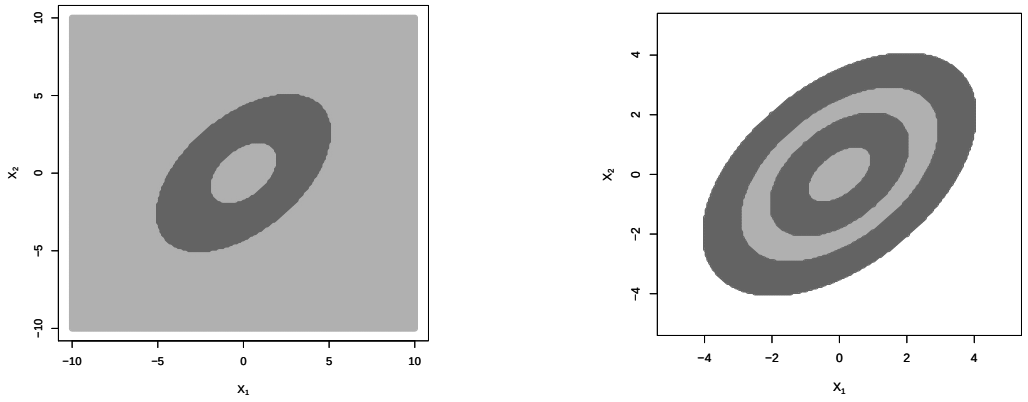
Keywords: Cluster analysis; cross-validation; elliptic distributions; EM algorithm; generalized additive models; logistic regression.

1. Introduction : history and motivation

More than seventy-five years ago, R. A. Fisher and P. C. Mahalanobis published their classic papers in the *Annals of Eugenics* (1936) and the *Proceedings of the National Institute of Sciences of India* (1936), respectively. It is probably not an overstatement that in several ways, these two papers have shaped the course of methodological and theoretical research in multivariate statistics for many future years. Mahalanobis (1936) defined a distance between two different populations as $(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$, which is popularly known as the Mahalanobis distance. Here μ_i is the mean of the i-th population for $i = 1, 2$, and Σ is their common dispersion matrix. His main idea was to make use of the correlations among different variables in the construction of this distance, and he had the multivariate normal distribution in his mind. On the other hand, Fisher (1936) was interested in the problem of discriminant analysis based on multivariate measurements for different populations. In particular, he considered the Iris data set that contains the four measurements : sepal length, sepal width, petal length and petal width of three different species (*Iris setosa*, *Iris versicolor* and *Iris virginica*) of the Iris flowers. He wanted to find a linear function of those four measurements that would maximize the "ratio of the difference between the specific means to the standard deviations within species" (Fisher (1936), p. 179). The linear function that maximizes this ratio is now popularly known as Fisher's linear

[a]**Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India.**
*__Email: tijahbus@gmail.com, subhajit_r@isical.ac.in__

discriminant function, and the maximum value of that ratio turns out to be $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, which is nothing but the Mahalanobis distance between the samples corresponding to two species. Here the multivariate measurements corresponding to the two species are denoted by $\mathbf{x}_1$ and $\mathbf{x}_2$; and $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and S denote the species means and the pooled dispersion matrix, respectively. Fisher's linear discriminant function classifies an observation $\mathbf{x}$ to the population that has the smallest Mahalanobis distance from that observation, where the Mahalanobis distance of an observation $\mathbf{x}$ and a population with mean μ and dispersion matrix Σ can be defined as $\mathrm{MD}(\mathbf{x}, \mu, \Sigma) = (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu)$. When different populations have normal distributions with a common scatter matrix and different means, the Bayes risk optimality of Fisher's linear discriminant function has been discussed in Welch (1939) and Rao (1948). Consider a two class problem with equal priors (i.e., $\pi_1 = \pi_2 = 1/2$) for the two classes, and the class densities f_1 and f_2 such that $f_i(\mathbf{x}) = |\Sigma|^{-1/2} g(\|\Sigma^{-1/2}(\mathbf{x} - \mu_i)\|)$ for $i = 1, 2$, where $g(\|\mathbf{x}\|)$ is a spherically symmetric unimodal density function with mode at the origin. Then, the Bayes rule will classify an observation $\mathbf{x}$ into the first class if and only if $\mathrm{MD}(\mathbf{x}, \mu_1, \Sigma) < \mathrm{MD}(\mathbf{x}, \mu_2, \Sigma)$. It is straight forward to see that this rule coincides with linear discriminant analysis (LDA). Besides, for two normal populations with different means (say, μ_1 and μ_2) and dispersions (say, Σ_1 and Σ_2), the Bayes classifier leads to quadratic discriminant analysis (QDA), where an observation $\mathbf{x}$ is classified into the first population if and only if $\mathrm{MD}(\mathbf{x}, \mu_1, \Sigma_1) - \mathrm{MD}(\mathbf{x}, \mu_2, \Sigma_2) < 2 \log(\pi_1/\pi_2) - \log(|\Sigma_1|/|\Sigma_2|)$. Here the left hand side of the inequality is a function of the difference between the Mahalanobis distances of the observation $\mathbf{x}$ from the two populations. Fisher's linear discriminant function and Mahalanobis distance are fundamental classification tools available in the literature, and both are widely used in practice.



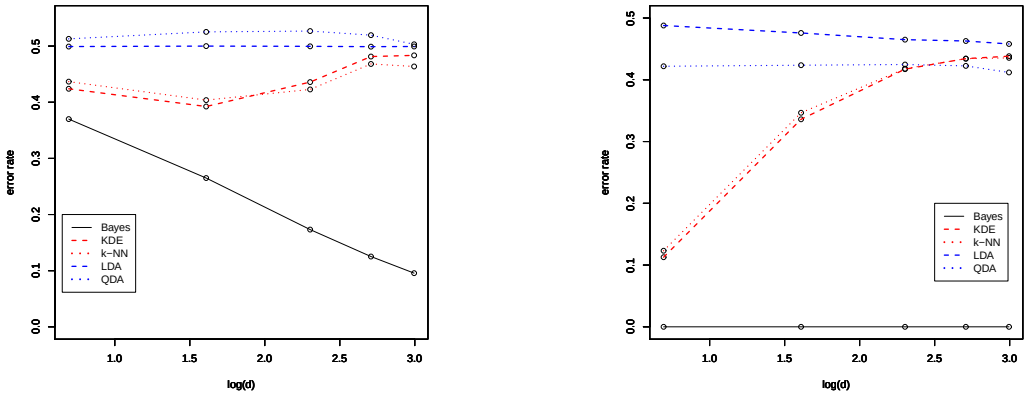
(a) **Example (a)** (b) **Example (b)**

Figure 1. Bayes class boundaries in two dimensions.

Consider now two examples denoted as **Example (a)** and **Example (b)**. **Example (a)** involves two classes in $\mathbb{R}^d$, where the distribution of the first class is an equal mixture of $\mathrm{N}_d(\mathbf{0}, \Sigma)$ and $\mathrm{N}_d(\mathbf{0}, 10\Sigma)$, and that for the second class is $\mathrm{N}_d(\mathbf{0}, 5\Sigma)$, where $\Sigma = [0.5\mathbf{1}'\mathbf{1} + 0.5\mathbf{I}_d]$. Here N_d denotes the d-dimensional multivariate normal distribution, $\mathbf{0}$ and $\mathbf{1}$ are the d-dimensional vectors of zeros and ones, respectively, and $\mathbf{I}_d$ is the $d \times d$ identity matrix. In **Example (b)**, each class distribution is an equal mixture of two uniform distributions. The first (respectively, the second) class distribution is a mixture of $\mathrm{U}_d(\mathbf{0}, \Sigma, 0, 1)$ and $\mathrm{U}_d(\mathbf{0}, \Sigma, 2, 3)$ (respectively, $\mathrm{U}_d(\mathbf{0}, \Sigma, 1, 2)$ and $\mathrm{U}_d(\mathbf{0}, \Sigma, 3, 4)$). Here $\mathrm{U}_d(\mu, \Sigma, r_1, r_2)$ denotes the uniform distribution over the region $\{\mathbf{x} \in \mathbb{R}^d : r_1 < \|\Sigma^{-1/2}(\mathbf{x} - \mu)\| < r_2\}$. In Figure 1., we show the class boundaries for the Bayes classifiers for these two examples in $\mathbb{R}^2$. The regions colored grey (respectively, black) correspond to observations being classified to the first (respectively, the second) class by the Bayes classifier. It is clear from the choice of these class distributions that LDA and QDA will not perform well in these examples. In fact, any classifier with linear or quadratic class boundaries (e.g., those based on support vector machines (SVM) with linear

or quadratic kernels) will deviate significantly from the Bayes classifiers in both the examples. In Figure 2., we plot misclassification rates of LDA and QDA along with the Bayes risks for **Examples (a)** and **(b)**, when $d = 2, 5, 10, 15$ and 20. The classifiers are trained on a sample of size 100 generated from each class, and the error rates are computed based on a sample of size 250 from each of the two classes. This procedure was repeated 500 times to calculate the average misclassification rates. Both LDA and QDA have very high misclassification rates that are sometimes close to 50%, and are much higher than the corresponding Bayes risks.

A natural question then is how some standard nonparametric classifiers like those based on k-NN (k-nearest neighbors) and KDE (kernel density estimates) (see, e.g., Duda, Hart and Stork (2000); Hastie *et. al.* (2009)) perform in such situations. In Figure 2., we plot the misclassification rates of these two classifiers as well. The smoothing parameters associated with the two classifiers (i.e., the k in k-NN and the bandwidth in KDE) were chosen by minimizing cross-validated estimates of the error rates (see, e.g., Hastie *et. al.* (2009)). For **Example (a)**, it is clear from Figure 2. that as the dimension increases, the Bayes risk decreases to zero, while in **Example (b)**, since the class distributions have disjoint supports, the Bayes risk is zero irrespective of the dimension of the data. However, the error rates for both the nonparametric classifiers increase to almost 50% in the two examples as the dimension increases.



(a) **Example (a)** (b) **Example (b)**

Figure 2. Misclassification rates of classifiers for $d = 2, 5, 10, 15$ and 20.

2. Classification using Mahalanobis distance

For the two examples in the preceding section, although the class distributions are elliptic, they have same location (i.e., $\mathbf{0}$), and they differ in their scatters as well as shapes. In **Example (a)**, the scatter matrices for the two classes are 5.5Σ and 5Σ, respectively. This implies that for any given observation, it will always be closer to the first population than the second in Mahalanobis distance. On the other hand, in **Example (b)**, one can show that the Mahalanobis distance of any observation $\mathbf{x}$ from the second population will be smaller than that from the first. This demonstrates that the simple classifier mentioned in Section 1, which is based on direct comparison of Mahalanobis distances, is not sufficient when the geometry of the class distributions is complex. We now consider situations where different populations are assumed to be elliptically symmetric but not necessarily unimodal. Further, the prior probabilities may be unequal, and the populations may differ in their scatters as well as shapes in addition to their locations.

2.1. Bayes classifier for elliptic populations

Let us assume that we have a J-class problem with $f_1, \cdots, f_J$ as the class densites, and the prior probabilities for the J populations are $\pi_1, \cdots, \pi_J$ with $\sum_{i=1}^{J} \pi_i = 1$. Suppose that f_i $(1 \leq i \leq J)$ is an elliptically symmetric density, where $f_i(\mathbf{x}) = |\Sigma_i|^{-1/2} g_i(\|\Sigma_i^{-1/2}(\mathbf{x} - \mu_i)\|)$ for $1 \leq i \leq J$, and $g_i(\|\mathbf{x}\|)$ is a spherically symmetric density on $\mathbb{R}^d$. Then the posterior probability for the i-th class $p(i|\mathbf{x})$ $(1 \leq i \leq J)$ is given by

$$p(i|\mathbf{x}) = \frac{\pi_i f_i(\mathbf{x})}{\sum_{k=1}^{J} \pi_k f_k(\mathbf{x})} = \frac{\pi_i |\Sigma_i|^{-1/2} g_i(\|\Sigma_i^{-1/2}(\mathbf{x} - \mu_i)\|)}{\sum_{k=1}^{J} \pi_k |\Sigma_k|^{-1/2} g_k(\|\Sigma_k^{-1/2}(\mathbf{x} - \mu_k)\|)}, \tag{1}$$

and the Bayes rule will assign an observation to the j-th class if $j = \arg\max_{1 \leq i \leq J} p(i|\mathbf{x})$. It follows from (1) that

$$\log\{p(i|\mathbf{x})/p(J|\mathbf{x})\} = \log(\pi_i |\Sigma_i|^{-1/2} / \pi_J |\Sigma_J|^{-1/2}) + \log g_i(\mathrm{MD}(\mathbf{x}, \mu_i, \Sigma_i)) - \log g_J(\mathrm{MD}(\mathbf{x}, \mu_J, \Sigma_J)),$$

for $1 \leq i \leq (J-1)$. This leads to the following result.

Result 1 : *When the class distributions are elliptically symmetric, the posterior probability $p(i|\mathbf{x})$ for the i-th class satisfies the multinomial additive logistic regression model given by*

$$p(i|\mathbf{x}) = p(i|\mathbf{z}(\mathbf{x})) = \frac{\exp(\Phi_i(\mathbf{z}(\mathbf{x})))}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}(\mathbf{x})))]} \tag{2}$$

for $1 \leq i \leq (J-1)$, and

$$p(J|\mathbf{x}) = p(J|\mathbf{z}(\mathbf{x})) = \frac{1}{[1 + \sum_{k=1}^{(J-1)} \exp(\Phi_k(\mathbf{z}(\mathbf{x})))]}, \tag{3}$$

where $\mathbf{z}(\mathbf{x}) = (MD(\mathbf{x}, \mu_1, \Sigma_1), MD(\mathbf{x}, \mu_2, \Sigma_2), \cdots, MD(\mathbf{x}, \mu_J, \Sigma_J))$, and $\Phi_k(\mathbf{z}) = \Phi_k(z_1, \cdots, z_J) = \sum_{m=1}^{J} \varphi_{km}(z_m)$ for $1 \leq k \leq (J-1)$. Here, for any $1 \leq i \leq (J-1)$, $\varphi_{ii}(z_i) = \log(\pi_i |\Sigma_i|^{-1/2}) + \log g_i(z_i)$, and $\varphi_{iJ}(z_J) = -\log(\pi_J |\Sigma_J|^{-1/2}) - \log g_J(z_J)$. Further, $\varphi_{ij}(z_j) = 0$ if $1 \leq i \neq j \leq (J-1)$.

In order to get some insights into the multinomial logistic regression model stated in Result 1, let us consider the case when each class distribution is normal. In such a situation, when the class dispersions are same, the logistic regression model in Result 1 becomes a linear logistic regression model related to LDA. On the other hand, when the class dispersions are different, we have a quadratic logistic regression model related to QDA. The implication of Result 1 is that when different class distributions are assumed to be elliptically symmetric, but nothing is assumed about the locations, the scatters and the shapes of the distributions, one still gets a multinomial additive logistic regression model with the Mahalanobis distances from different populations as the vector of co-variates.

The expressions for class posteriors in (2) and (3) lead to a semiparametric approach for constructing the classifier with elliptic class distributions. We can begin by constructing some suitable estimates of the locations and the scatters for different classes, and compute the Mahalanobis distances of an observation from each of the J-classes. Then the J-dimensional vector of co-variates can be formed with those distances as its co-ordinates. Unlike standard nonparametric classification tools, this automatically leads to a substantial reduction in data dimension in many situations in practice. Further, since we have an additive logistic regression model in Result 1, it guards against the curse of dimensionality in the estimation of the class posteriors even if J is large. On one hand, a classifier based on estimates of class posteriors given in (2) and (3) will be more flexible than standard parametric classifiers having linear or quadratic class boundaries - on the other hand, it will not suffer from the curse of dimensionality that affects standard nonparametric classifiers.

In the machine learning literature, there has been some work on modifying LDA to yield non-linear decision boundaries. For instance, Mika *et. al.* (1999) modified LDA and proposed a "kernelized" version of Fisher's LDA. A notion of

"local LDA", which aims at capturing multimodality in a dataset, has been developed in the paper by Sugiyama (2007). However, it is not clear how such classifiers relate to the Bayes classifier for appropriate class distributions, and no theoretical result is known about their misclassification probabilities. In the next subsection, we develop a semiparametric classifier that can handle multimodal class distributions, which are modelled as mixtures of elliptic distributions.

2.2. Classification using mixture models

Hastie and Tibshirani (1996) proposed a generalization of LDA by modelling the density function of each class by a finite mixture of normal densities, and they called their method mixture discriminant analysis (MDA). Later, Fraley and Raftery (2002) extended MDA to MclustDA, where the number of components in the mixture is chosen using the Bayes information criterion (BIC). In the implementation of both MDA and MclustDA, the class distributions are assumed to be mixtures of normal distributions, and that may not be justified in many situations in practice. For a detailed discussion of several mixture models with non-normal components and their applications, one may see the book by McLachlan and Peel (2000). We now consider the situation where the class distributions are mixtures of finitely many elliptic distributions with arbitrary locations, scatters and shapes. This provides us with a very rich class of models.

Assume that the density function of the i-th class is a finite mixture of elliptically symmetric densities that can be written as $f_i(\mathbf{x}) = \sum_{k=1}^{R_i} \theta_{ik} |\Sigma_{ik}|^{-1/2} g_{ik}(\|\Sigma_{ik}^{-1/2}(\mathbf{x} - \mu_{ik})\|)$, where θ_{ik}s are positive satisfying $\sum_{k=1}^{R_i} \theta_{ik} = 1$ for all $1 \leq i \leq J$. One can interpret an observation arising from such a mixture distribution as an observation from one of the R_i sub-populations, where the sub-population is chosen randomly in such a way that the k-th sub-population ($1 \leq k \leq R_i$) has probability θ_{ik} of being selected. Note that for any $1 \leq i \leq J$, the k-th sub-class in the i-th class has elliptic density g_{ik} with location parameter μ_{ik} and scatter matrix Σ_{ik} for $1 \leq k \leq R_i$. In this case, given an observation $\mathbf{x}$, we have the following result for the conditional probability that it belongs to the i-th class.

Result 2 : *When the class distributions are mixtures of elliptically symmetric distributions, the posterior probability for the i-th class is $p(i|\mathbf{x}) = \sum_{r=1}^{R_i} p(c_{ir}|\mathbf{x})$ for all $1 \leq i \leq J$. Here, the posterior probability $p(c_{ir}|\mathbf{x})$ for the r-th sub-class in the i-th class, which is denoted by c_{ir}, satisfies the multinomial additive logistic regression model given by*

$$p(c_{ir}|\mathbf{x}) = p(c_{ir}|\mathbf{z}(\mathbf{x})) = \frac{\exp(\Phi_{ir}(\mathbf{z}(\mathbf{x})))}{[1 + \sum_{m=1}^{J} \sum_{k=1, \{m \neq J,\ k \neq R_J\}}^{R_m} \exp(\Phi_{mk}(\mathbf{z}(\mathbf{x})))]}, \tag{4}$$

for $1 \leq r \leq R_i$ and $1 \leq i \leq J$ with $(i, r) \neq (J, R_J)$, and

$$p(c_{JR_J}|\mathbf{x}) = p(c_{JR_J}|\mathbf{z}(\mathbf{x})) = \frac{1}{[1 + \sum_{m=1}^{J} \sum_{k=1, \{m \neq J,\ k \neq R_J\}}^{R_m} \exp(\Phi_{mk}(\mathbf{z}(\mathbf{x})))]}. \tag{5}$$

Here $\mathbf{z}(\mathbf{x}) = (MD(\mathbf{x}, \mu_{11}, \Sigma_{11}), \cdots, MD(\mathbf{x}, \mu_{1R_1}, \Sigma_{1R_1}), \cdots, MD(\mathbf{x}, \mu_{J1}, \Sigma_{J1}), \cdots, MD(\mathbf{x}, \mu_{JR_J}, \Sigma_{JR_J}))$ is the vector of co-variates, and $\Phi_{ir}(\mathbf{z}) = \Phi_{ir}(z_{11}, \cdots, z_{1R_1}, \cdots, z_{J1}, \cdots, z_{JR_J}) = \sum_{m=1}^{J} \sum_{k=1}^{R_m} \varphi_{ir,mk}(z_{mk})$ for $1 \leq r \leq R_i$ and $1 \leq i \leq J$. Further, for any $1 \leq r \leq R_i$ and $1 \leq i \leq J$, $\varphi_{ir,ir}(z_{ir}) = \log(\pi_i \theta_{ir} |\Sigma_{ir}|^{-1/2}) + \log g_{ir}(z_{ir})$ except for φ_{JR_J,JR_J}, and $\varphi_{JR_J,JR_J}(z_{JR_J}) = -\log(\pi_J \theta_{JR_J} |\Sigma_{JR_J}|^{-1/2}) - \log g_{JR_J}(z_{JR_J})$. Also, $\varphi_{ir,i'r'}(z_{i'r'}) = 0$ whenever $1 \leq r \neq r' \leq R_i$ or $1 \leq i \neq i' \leq J$.

3. SPARC : a semiparametric classification procedure

We now describe a semiparametric classifier developed using the results derived in the preceding section. Suppose that we have the training observations $(\mathbf{x}_l, y_l) \in \mathbb{R}^d \times \{1, \cdots, J\}$, where y_l denotes the class label for the l-th observation for $1 \leq l \leq n$. The estimation of the posterior probabilities $p(c_{ir}|\mathbf{x})$'s in (4) and (5) involves the estimation of the

location parameters μ_{ir}s, the scatter matrices Σ_{ir}s and the unknown functions Φ_{ir}s for $1 \leq r \leq R_i$ and $1 \leq i \leq J$. If we had known the sub-class label of each observation, these objects could have been estimated as follows. The estimates of μ_{ir}s and Σ_{ir}s would have been

$$\hat{\mu}_{ir} = \frac{\sum_{l:y_l=i} \mathbf{x}_l I(\mathbf{x}_l \in c_{ir})}{\sum_{l:y_l=i} I(\mathbf{x}_l \in c_{ir})} \text{ and } \hat{\Sigma}_{ir} = \frac{\sum_{l:y_l=i} (\mathbf{x}_l - \hat{\mu}_{ir})(\mathbf{x}_l - \hat{\mu}_{ir})' I(\mathbf{x}_l \in c_{ir})}{\sum_{l:y_l=i} I(\mathbf{x}_l \in c_{ir})}, \tag{6}$$

where $I(\cdot)$ is the usual indicator function. Then, in view of Result 2, we could estimate $p(c_{ir}|\mathbf{x})$'s by fitting a multinomial nonparametric additive logistic regression using the training data, where we could use

$$\hat{\mathbf{z}}(\mathbf{x}_l) = (\text{MD}(\mathbf{x}_l, \hat{\mu}_{11}, \hat{\Sigma}_{11}), \cdots, \text{MD}(\mathbf{x}_l, \hat{\mu}_{1R_1}, \hat{\Sigma}_{1R_1}), \cdots, \text{MD}(\mathbf{x}_l, \hat{\mu}_{J1}, \hat{\Sigma}_{J1}), \cdots, \text{MD}(\mathbf{x}_l, \hat{\mu}_{JR_J}, \hat{\Sigma}_{JR_J})) \tag{7}$$

as the vector of co-variates, and

$$\mathbf{I}(\mathbf{x}_l) = (I(\mathbf{x}_l \in c_{11}), \cdots, I(\mathbf{x}_l \in c_{1R_1}), \cdots, I(\mathbf{x}_l \in c_{J1}), \cdots, I(\mathbf{x}_l \in c_{JR_J})) \tag{8}$$

as the response vector for $1 \leq l \leq n$. For fitting the multinomial nonparametric additive logistic regression model, the backfitting algorithm of Hastie and Tibshirani (1990) can be used. Since the $\varphi_{ir,mk}$s for $1 \leq r, k \leq R_i$ and $1 \leq i, m \leq J$ satisfy certain constraints stated in Result 2, one has to solve a constrained optimization problem for estimating the conditional sub-class probabilities. Programs imposing such constraints on the functions in generalized additive models (see Yee and Wild (1996)) are available in the R library `VGAM`.

Since the sub-class labels are actually not available in the training data, we need to treat those labels as missing data (cf. Hastie and Tibshirani (1996); Fraley and Raftery (2002)). We can use an 'EM type' (see, e.g., Dempster, Laird and Rubin (1977); McLachlan and Krishnan (1997)) iterative procedure for constructing the estimates of posterior probabilities $p(c_{ir}|\mathbf{x})$s. The 'E-step' of that iterative procedure will involve construction of estimates of the response vector $\mathbf{I}(\mathbf{x})$, while the 'M-step' will involve the estimation of μ_{ir}s, Σ_{ir}s and $p(c_{ir}|\mathbf{x})$s for $1 \leq r \leq R_i$ and $1 \leq i \leq J$.

If we assume that the number of sub-classes (i.e., the R_is) are known, we can use an appropriate clustering technique to form the sub-classes within a class using the training data. This will provide the initial estimate for the response vectors $\hat{\mathbf{I}}^{(0)}(\mathbf{x}_l) = (\hat{I}(\mathbf{x}_l \in c_{11}), \cdots, \hat{I}(\mathbf{x}_l \in c_{1R_1}), \cdots, \hat{I}(\mathbf{x}_l \in c_{J1}), \cdots, \hat{I}(\mathbf{x}_l \in c_{JR_J}))$ for $1 \leq l \leq n$, where the estimate of the sub-class label of $\mathbf{x}_l$ is determined by the cluster to which it belongs. Then, the initial estimates of μ_{ir} and Σ_{ir}, say $\hat{\mu}_{ir}^{(0)}$ and $\hat{\Sigma}_{ir}^{(0)}$, can be obtained by substituting $\hat{I}^{(0)}(\mathbf{x}_l \in c_{ir})$ in place of $I(\mathbf{x}_l \in c_{ir})$ in (6). The co-variate vector in (7) can be modified by replacing μ_{ir} and Σ_{ir} by $\hat{\mu}_{ir}^{(0)}$ and $\hat{\Sigma}_{ir}^{(0)}$, respectively, where $1 \leq r \leq R_i$ and $1 \leq i \leq J$. We denote this co-variate vector as $\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)$. A similar approach has been used earlier by Hastie and Tibshirani (1996); Fraley and Raftery (2002); Celeux and Govaert (1992), who considered finite mixtures of Gaussian or other parametric families of distributions. These authors formed estimates for the class posteriors by plugging in $\hat{\mu}_{ir}^{(0)}$ and $\hat{\Sigma}_{ir}^{(0)}$ into the expressions of those posterior probabilities derived from the parametric models they used. However, since we do not assume any parametric model for the components of the mixture distribution, we need to carry out an additional estimation step in our 'M-step' to estimate the conditional class probabilities. The initial estimates for the conditional sub-class probabilities $\hat{p}^{(0)}(c_{ir}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l))$ for $1 \leq r \leq R_i$ and $1 \leq i \leq J$ can be computed by fitting a constrained multinomial nonparametric additive logistic regression model, which is described in the preceding paragraph. Here, we use $\hat{\mathbf{I}}^{(0)}(\mathbf{x}_l)$ as the response vectors and $\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)$ as the co-variate vectors for $1 \leq l \leq n$.

In the next iteration, if $\mathbf{x}_l$ belongs to the s-th class, i.e., $\mathbf{x}_l \in \cup_{r=1}^{R_s} c_{sr}$, we can implement the 'E-step' by forming

$$\hat{\mathbf{I}}^{(1)}(\mathbf{x}_l) = (\hat{q}^{(0)}(c_{11}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)), \cdots, \hat{q}^{(0)}(c_{1R_1}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)), \cdots, \hat{q}^{(0)}(c_{J1}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)), \cdots, \hat{q}^{(0)}(c_{JR_J}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l))),$$

where $\hat{q}^{(0)}(c_{sr}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)) = \hat{p}^{(0)}(c_{sr}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)) / \sum_{v=1}^{R_s} \hat{p}^{(0)}(c_{sv}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l))$ for $1 \leq r \leq R_s$, and $\hat{q}^{(0)}(c_{ir}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l)) = 0$ for all

$i \neq s$. Note that the construction of $\hat{q}^{(0)}(c_{ir}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l))$ uses the class label of $\mathbf{x}_l$, which is known though the sub-class label of $\mathbf{x}_l$ is not known. Then, for the next 'M-step', we compute $\hat{\mu}_{ir}^{(1)}$ and $\hat{\Sigma}_{ir}^{(1)}$ by replacing the indicators in (6) by the corresponding components of $\hat{\mathbf{l}}^{(1)}(\mathbf{x}_l)$. Using (7), we obtain the co-variate vector $\hat{\mathbf{z}}^{(1)}(\mathbf{x}_l)$, which is now based on $\mu_{ir}^{(1)}$ and $\Sigma_{ir}^{(1)}$. For $1 \leq l \leq n$, in order to obtain the estimates $\hat{p}^{(1)}(c_{ir}|\hat{\mathbf{z}}^{(0)}(\mathbf{x}_l))$ for $1 \leq r \leq R_i$ and $1 \leq i \leq J$, we again need to solve the problem of fitting a nonparametric additive multinomial logistic regression model using the co-variate vector $\hat{\mathbf{z}}^{(1)}(\mathbf{x}_l)$, and the non-standard version of the response vector $\hat{\mathbf{l}}^{(1)}(\mathbf{x}_l)$. Since the components of $\hat{\mathbf{l}}^{(1)}(\mathbf{x}_l)$ are no longer guaranteed to be 0s and 1s only, we need a modified version of the constrained backfitting algorithm (see Yee and Wild (1996)). We have appropriately modified the R codes of `VGAM` to cope with this non-standard form of the response vector. We may continue the iterations to generate the sequence of estimates of the class posteriors until the estimates arising from two consecutive iterations are sufficiently close. If we terminate after the K-th iteration, the final posterior estimates of the i-th class for a new observation $\mathbf{x}$ becomes $\hat{p}(i|\mathbf{x}) = \sum_{r=1}^{R_i} \hat{p}^{(K)}(c_{ir}|\hat{\mathbf{z}}^{(K)}(\mathbf{x}))$ for $1 \leq i \leq J$, and $\mathbf{x}$ is classified to the j-th class if $j = \arg\max_{1 \leq i \leq J} \hat{p}(i|\mathbf{x})$. We call this classification procedure *SPARC*, which is an acronym obtained by abbreviating *SemiPARametric Classification*.

Given the values of R_is, there are many clustering algorithms available in the literature (see, e.g., Duda, Hart and Stork (2000); Everitt, Landau and Lesse (2001)) that can be used for the cluster analysis of each class and to form the initial estimates of the response vectors $\hat{\mathbf{l}}^{(0)}(\mathbf{x}_l)$ for $1 \leq l \leq n$. In SPARC, we have used the well-known k-means algorithm, which is computationally very simple and produced good results in our empirical study. Hastie and Tibshirani (1996) also used the k-means algorithm for initial clustering of the data in MDA.

In MDA (Hastie and Tibshirani (1996)), the values of R_is are specified. The default value is $R_i = 3$ for all i in the R code `mda`. However, in MclustDA, Fraley and Raftery (2002) proposed to use BIC to estimate the values of R_is. This algorithm uses a default upper bound of 9 in the R code `mclustDAtrain`. We have observed in our numerical investigation that BIC often leads to over-estimation, and this results in a large number of mixing components for each class distribution. Consequently, in several data sets analyzed in sub-section 4.2, MclustDA had much higher misclassification rates compared to other classifiers. In SPARC, we are not assuming Gaussian distribution or any parametric model for the components of the mixture distributions, and we have decided to choose the values of R_is by minimizing a 5-fold cross-validated estimate (see, e.g., Hastie *et. al.* (2009) for a discussion on V-fold cross-validation) of the misclassification rates. To implement the cross-validation in SPARC, we needed some appropriate upper bounds for the R_is, and we have used the 'Gap Statistic' (see Tibshirani, Walther and Hastie (2001)) alongwith the k-means algorithm to obtain this upper bound for each class.

4. Comparison of different classifiers

In this section, we compare SPARC with a pool of other well-known classifiers when applied to several simulated and real data sets. The books by Duda, Hart and Stork (2000) and Hastie *et. al.* (2009) contain detailed descriptions of these classifiers and discussion about their performance. In addition to MDA and MclustDA, this pool of classifiers includes standard parametric classifiers like LDA and QDA, as well as standard nonparametric classifiers based on k-NN and KDE. For both KDE and k-NN, we have standardized the data vectors using the variance covariance matrix. However, for KDE we have used separate scatter matrices for different classes, while for k-NN, we have used the pooled scatter matrix as is usually done in practice. The parameters k in k-NN and the bandwidth in KDE were chosen by minimizing a cross-validated estimate of their misclassification rates. We also study the performance of SPARC in comparison with that of SVM with the linear kernel and the radial basis function with default values of the associated parameters given in the R codes in http://www.csie.ntu.edu.tw/∼cjlin/libsvm/, and the classifiers based on CART (classification and regression trees) and Poly-MARS (polychotomous regression with multivariate adaptive regression splines). For implementation of SVM, CART and Poly-MARS, we have used R codes available in the libraries `e1071`, `tree` and `mars`, respectively. We have written our own R codes for SPARC (which is available at

www.isical.ac.in/~tijahbus/SC.htm), and our iterations were terminated at $K = 2$ in most cases.

4.1. Analysis of simulated data

In addition to the two examples (i.e., **Examples (a)** and **(b)**) introduced in Section 1, we consider three simulated examples in this section, and each of these is formed with two classes. In the third example, i.e., **Example (c)**, we choose the distributions of the two classes to be $t_{3,d}(\mathbf{0}, \mathbf{I}_d)$ and $t_{3,d}(\mathbf{1}, 4\mathbf{I}_d)$, where $t_{3,d}$ denotes the d-dimensional t distribution with 3 d.f. The next two examples involve mixtures of elliptic distributions. In **Example (d)**, the first class is an equal mixture of $N_d(\mathbf{0}, 0.25\mathbf{I}_d)$ and $N_d(\mathbf{0}, \mathbf{I}_d)$, and the second class is an equal mixture of $N_d(\text{-}\mathbf{1}, 0.25\mathbf{I}_d)$ and $N_d(\mathbf{1}, 0.25\mathbf{I}_d)$, where N_d is the d-variate normal distribution as in Section 1. On the other hand, in **Example (e)**, one class distribution is an equal mixture of $U_d(\mathbf{0}, \Sigma, 0, 1)$ and $U_d(\mathbf{2}, \Sigma, 2, 3)$, and the other one is an equal mixture of $U_d(\mathbf{1}, \Sigma, 1, 2)$ and $U_d(\mathbf{3}, \Sigma, 3, 4)$, where $U_d(\mu, \Sigma, r_1, r_2)$ is the uniform distribution as defined in Section 1. Here $\mathbf{0}$, $\mathbf{1}$, $\mathbf{2}$, $\mathbf{3}$ are d-dimensional vectors of zeros, ones, twos and threes, respectively, and $\mathbf{I}_d$ is the $d \times d$ identity matrix. In each example, taking an equal number of observations from each of the two classes, we generated 500 training and test sets. Each training set consists of 200 (=100+100 for the two classes) observations, and each test set contains 500 (=250+250 for the two classes) observations. Average test set misclassification rates (over 500 trials) of different classifiers are reported in Table 1 along with their corresponding standard errors. We consider **Examples (a)** - **(e)** in dimensions 5, 10 and 20, and to facilitate comparison, we also report the corresponding Bayes risk in each case.

In **Examples (a)** and **(b)**, SPARC outperforms all other competing classifiers in all dimensions. In **Example (c)**, where the class distributions are $t_{3,d}$, SVM with radial basis function yields the best performance with SPARC having the third best performance. In **Example (d)** involving bimodal distributions, the classifier based on MclustDA turns out to have the best error rate with SPARC having the second best error rate. In this example, MDA fails to yield satisfactory performance, and the difference between the performance of SPARC and MDA increases as the dimension grows. In **Example (e)**, which involves class distributions that are mixtures of non-Gaussian distributions, SPARC significantly outperforms all classifiers.

Classifiers like LDA, QDA and SVM with linear kernel fail in examples where the geometry of the underlying distributions is complex. However, SVM with radial basis function is more flexible, and it yields significant improvement in such examples. The performance of nonparametric classifiers based on k-NN and KDE deteriorates as the data dimension increases. While CART does not seem to be severely affected by the increase in data dimension, it also fails to capture the optimal class boundaries when the geometry of the class distributions is complex. Like CART, Poly-MARS also uses adaptive partitioning of the co-variate space, but it constructs the classifier by formulating the classification problem as a polychotomous logistic regression problem, where the regression model involves additive as well as first order interaction terms. Although Poly-MARS improves over CART in the examples considered here, it is far from being optimal. MDA assumes a common covariance matrix over all the classes, and hence performs badly when the underlying populations are mixtures of normal distributions, which differ in their scatters. Unlike MDA, both MclustDA and SPARC allow for unequal estimates for the scatter matrices in each sub-class. However, we have observed that if we pool different sub-class estimates within a class, the performance of SPARC improves in our examples, and we have implemented it accordingly. Both MDA and MclustDA perform poorly when the class distributions are mixtures of non-Gaussian distributions.

4.2. Analysis of real benchmark data sets

We now analyze some real benchmark data sets. Among these data sets, the hemophilia data is obtained from Johnson and Wichern (1992). All other data sets are taken either from the UCI machine learning repository (http://archive.ics.uci.edu/ml/) or from the CMU data archive (http://lib.stat.cmu.edu/datasets/). Descriptions of the data sets are available at these sources. For the biomedical data, we did not consider the observations with missing values. Each data set was randomly partitioned 500 times into training and test sets (see Table 2 for their sizes).

Table 1. Misclassification rates (in %) of different classifiers and their standard errors (in parantheses) for simulated data sets. For each example, the minimum misclassification rate of a classifier is written in bold.

Data set	Bayes risk	LDA	QDA	SVM (linear)	SVM (radial)	k-NN	KDE	CART	Poly -MARS	MDA	Mclust -DA	SPARC
$d = 5$												
Ex (a)	26.50	50.00	52.53	45.42	30.67	40.38	39.24	39.39	39.31	38.97	30.93	**29.30**
		(0.10)	(0.19)	(0.11)	(0.09)	(0.11)	(0.12)	(0.14)	(0.20)	(0.11)	(0.10)	(0.10)
Ex (b)	0	47.58	42.36	44.02	37.70	34.66	33.60	39.54	42.35	41.75	31.46	**9.69**
		(0.15)	(0.06)	(0.10)	(0.09)	(0.14)	(0.11)	(0.13)	(0.21)	(0.09)	(0.13)	(0.10)
Ex (c)	18.10	25.57	24.96	26.86	**19.76**	24.17	25.15	31.53	26.83	24.89	20.19	20.60
		(0.11)	(0.13)	(0.11)	(0.08)	(0.10)	(0.11)	(0.13)	(0.15)	(0.10)	(0.09)	(0.10)
Ex (d)	4.05	49.73	16.62	47.77	6.03	8.98	8.30	21.81	17.27	6.60	**4.33**	6.06
		(0.10)	(0.11)	(0.13)	(0.04)	(0.08)	(0.08)	(0.12)	(0.16)	(0.05)	(0.04)	(0.07)
Ex (e)	5.60	40.26	44.24	42.88	24.58	26.57	25.72	31.61	31.48	30.08	23.07	**14.28**
		(0.08)	(0.06)	(0.07)	(0.10)	(0.11)	(0.10)	(0.12)	(0.11)	(0.10)	(0.12)	(0.10)
$d = 10$												
Ex (a)	17.32	49.96	52.68	45.50	28.47	42.28	43.59	39.23	38.67	39.51	29.44	**22.20**
		(0.10)	(0.21)	(0.11)	(0.08)	(0.13)	(0.13)	(0.14)	(0.17)	(0.11)	(0.10)	(0.11)
Ex (b)	0	46.50	42.47	44.50	37.42	41.80	41.72	40.35	42.68	42.34	25.03	**12.47**
		(0.14)	(0.06)	(0.13)	(0.08)	(0.09)	(0.08)	(0.13)	(0.21)	(0.10)	(0.13)	(0.10)
Ex (c)	12.58	20.07	20.67	21.11	**14.31**	21.98	22.55	31.88	23.90	20.06	15.80	17.25
		(0.10)	(0.12)	(0.10)	(0.07)	(0.11)	(0.10)	(0.15)	(0.15)	(0.10)	(0.10)	(0.10)
Ex (d)	0.50	49.24	14.65	47.96	1.87	17.25	15.34	21.77	15.69	2.43	**1.26**	1.77
		(0.11)	(0.07)	(0.11)	(0.02)	(0.11)	(0.10)	(0.12)	(0.16)	(0.05)	(0.05)	(0.05)
Ex (e)	0.75	41.26	43.97	42.37	19.93	35.09	34.86	26.85	25.87	23.95	15.25	**12.12**
		(0.07)	(0.05)	(0.07)	(0.08)	(0.13)	(0.11)	(0.11)	(0.10)	(0.09)	(0.11)	(0.10)
$d = 20$												
Ex (a)	9.56	49.93	50.32	45.49	26.80	46.37	48.35	39.16	37.91	40.37	28.51	**21.29**
		(0.10)	(0.22)	(0.10)	(0.08)	(0.12)	(0.11)	(0.14)	(0.15)	(0.11)	(0.10)	(0.15)
Ex (b)	0	45.80	41.19	44.54	37.32	43.55	43.82	40.40	42.77	42.46	22.96	**20.48**
		(0.13)	(0.06)	(0.12)	(0.08)	(0.09)	(0.09)	(0.13)	(0.20)	(0.09)	(0.12)	(0.10)
Ex (c)	7.82	14.49	16.31	15.29	**9.32**	20.05	20.53	31.88	14.82	20.34	14.67	15.18
		(0.09)	(0.09)	(0.09)	(0.06)	(0.11)	(0.11)	(0.15)	(0.09)	(0.13)	(0.10)	(0.10)
Ex (d)	0.12	49.56	15.46	48.14	0.64	33.64	33.39	21.75	14.37	0.73	**0.24**	0.44
		(0.10)	(0.07)	(0.10)	(0.05)	(0.11)	(0.10)	(0.12)	(0.18)	(00.02)	(0.02)	(0.03)
Ex (e)	0	40.97	42.28	42.32	11.69	44.88	43.52	18.58	16.72	15.29	12.08	**10.43**
		(0.08)	(0.06)	(0.07)	(0.08)	(0.09)	(0.10)	(0.11)	(0.09)	(0.08)	(0.09)	(0.08)

Table 2. Misclassification rates (in %) of different classifiers and their standard errors (in parantheses) for real data sets. For each data set, the minimum misclassification rate of a classifier is written in bold.

Data (d, J)	Train size	Test size	LDA	QDA	SVM (linear)	SVM (radial)	k-NN	KDE	CART	Poly -MARS	MDA	Mclust -DA	SPARC
Hemophilia	50	25	15.22	15.47	16.78	16.02	15.79	15.11	19.10	15.26	16.18	22.64	**14.54**
(2,2)			(0.27)	(0.26)	(0.28)	(0.28)	(0.30)	(0.27)	(0.30)	(0.28)	(0.29)	(0.43)	(0.33)
Biomed	100	94	15.66	12.57	22.03	12.76	18.00	16.93	19.07	14.91	12.97	19.54	**12.15**
(4,2)			(0.14)	(0.12)	(0.18)	(0.12)	(0.15)	(0.15)	(0.17)	(0.17)	(0.13)	(0.41)	(0.13)
Pima	384	384	**23.37**	25.99	23.67	24.19	25.73	26.57	27.20	23.84	24.68	32.67	25.23
(8,2)			(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)	(0.09)	(0.07)	(0.08)	(0.19)	(0.08)
Iris	90	60	**2.30**	2.68	3.46	4.11	2.45	2.49	6.22	5.10	2.88	14.81	5.29
(4,3)			(0.15)	(0.17)	(0.09)	(0.10)	(0.07)	(0.07)	(0.11)	(0.10)	(0.08)	(0.60)	(0.17)
Blood	100	45	10.46	9.39	9.93	15.76	10.11	11.16	**2.46**	3.12	7.77	23.11	9.32
(5,3)			(0.18)	(0.18)	(0.20)	(0.21)	(0.18)	(0.19)	(0.11)	(0.13)	(0.17)	(0.39)	(0.19)
Wine	100	78	2.00	2.46	3.64	1.86	2.04	**1.66**	10.22	6.41	2.06	8.00	2.58
(13,3)			(0.06)	(0.09)	(0.09)	(0.06)	(0.07)	(0.06)	(0.19)	(0.14)	(0.07)	(0.52)	(0.08)
Vehicle	422	424	22.49	**16.38**	21.20	25.57	21.84	21.45	42.12	23.76	20.79	18.64	17.09
(18,4)			(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)	(0.12)	(0.11)	(0.08)	(0.13)	(0.10)

For the hemophilia data as well as the biomedical data, SPARC yields the best error rate and outperforms all other classifiers that we have used in the comparison. In the vehicle data set, QDA yields the lowest misclassification rate, and the misclassification rate of SPARC is the second best. LDA yields the best error rate for the pima data set, and SPARC performs worse than five of the other classifiers and better than the remaining five. Interestingly, in the case of the famous iris data which was analyzed by Fisher (1936), standard parametric classifiers like LDA and QDA as well as well-known nonparametric classifiers like k-NN and KDE perform better than other classifiers. CART and Poly-MARS perform significantly better than all other classifiers in the case of blood data. For this data set, only three of the classifiers perform better than SPARC. In the wine data, the data dimension is 13, and the sizes of the training sets were 33, 40 and 27. As a consequence, separate estimates of class dispersions suffer from statistical instability. The use of a pooled estimate of the scatter matrix significantly improved the misclassification rates for KDE and SPARC. The classifier based on KDE performs best for this data set, and SPARC yields a reasonable misclassification rate, which is higher than that of six of the other classifiers and lower than that of the remaining four.

Interestingly, except for the vehicle data set, MDA performs better than MclustDA in all data sets. In fact, MclustDA yields the worst error rate in four data sets. We have observed that the use of BIC in choosing the number of components in the mixture of Gaussian distributions leads to over-estimation of the number of sub-classes, which affects its performance. In SPARC, we minimize the cross-validated estimate of the misclassification rates to estimate the number of sub-classes, and this leads to a significantly improved performance. In the case of vehicle data, QDA has the best performance, and the performance of LDA is much worse, which indicates that different class distributions differ mainly in their scatters, and not so much in their locations. For this data set, the performance of MDA, which uses a common estimate of the scatter matrix for all the classes, is significantly worse than that of MclustDA. In the wine and the vehicle data sets, CART has the worst performance, while SVM with linear kernel has worst performance for the biomedical data.

References

Celeux, G. and Govaert, G. (1992), 'A classification EM algorithm for clustering and two stochastic versions', *Comp. Statist. Data Anal.*, **14**, 315–332.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood for incomplete data via the EM algorithm', *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Duda, R., Hart, P. and Stork, D. G. (2000), *Pattern Classification*, Wiley, New York.

Everitt, B. S., Landau, S. and Lesse, M. (2001), *Cluster Analysis*, Arnold, London.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Ann. Eugenics*, **7**, 179–188.

Fraley, C. and Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *J. Amer. Statist. Assoc.*, **97**, 611–631.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, New York.

Hastie, T. and Tibshirani, R. (1996), 'Discriminant analysis by Gaussian mixtures', *J. Roy. Statist. Soc. Ser. B*, **58**, 155–176.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *Elements of Statistical Learning Theory*, Springer, New York.

Johnson, R. A. and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, Prentice-Hall.

Mahalanobis, P. C. (1936), 'On the generalized distance in statistics', *Proc. Nat. Acad. Sci., India*, **12**, 49–55.

McLachlan, G. J. and Krishnan, T. (1997), *The EM algorithm and Extensions*, Wiley, New York.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Muller, K.-R. (1999), 'Fisher discriminant analysis with kernels', In Y.-H. Hu, J. Larsen, E. Wilson and S. Douglas, ed. *Neural Networks for Signal Processing IX*, 41-48.

Rao, C. R. (1948), 'The utilization of multiple measurements in problems of biological classification', *J. Roy. Statist. Soc. Ser. B*, **10**, 159–203.

Sugiyama, M. (2007), 'Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis', *J. Mach. Learn. Res.*, **8**, 1027–1061.

Tibshirani, R., Walther, G. and Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *J. Roy. Statist. Soc. Ser. B*, **63**, 411–423.

Welch, B. L. (1939), 'Note on discriminant functions', *Biometrika*, **31**, 218–220.

Yee, T. W. and Wild, C. J. (1936), 'Vector generalized additive models', *J. Roy. Statist. Soc. Ser. B*, **58**, 481–493.