Comput Stat (2010) 25:121–141
DOI 10.1007/s00180-009-0165-9

ORIGINAL PAPER

Penalized multinomial mixture logit model

Shaheena Bashir · Edward M. Carter

Received: 24 June 2008 / Accepted: 27 July 2009 / Published online: 14 August 2009

Abstract Normal distribution based discriminant methods have been used for the classification of new entities into different groups based on a discriminant rule constructed from the learning set. In practice if the groups are not homogeneous, then mixture discriminant analysis of Hastie and Tibshirani (J R Stat Soc Ser B 58(1):155–176, 1996) is a useful approach, assuming that the distribution of the feature vectors is a mixture of multivariate normals. In this paper a new logistic regression model for heterogenous group structure of the learning set is proposed based on penalized multinomial mixture logit models. This approach is shown through simulation studies to be more effective. The results were compared with the standard mixture discriminant analysis approach using the probability of misclassification criterion. This comparison showed a slight reduction in the average probability of misclassification using this penalized multinomial mixture logit model as compared to the classical discriminant rules. It also showed better results when applied to practical life data problems producing smaller errors.

Keywords Mixture models · EM algorithm · Logit models · Penalty parameter · Leukemia data

S. Bashir (✉)
Department of Mathematics and Statistics, McMaster University,
Hamilton, ON L8S 4L8, Canada
e-mail: sbashir@math.mcmaster.ca

E. M. Carter
Department of Mathematics and Statistics, University of Guelph, Guelph,
ON N1G 2W1, Canada

1 Introduction

Consider the usual classification situation: the training sample consists of class membership matrix $Y_{n\times J}$ for $i = 1, \ldots, n$ observations, $j = 1, \ldots, J$ groups and the predictor matrix $X_{n\times p}$. The goal is to predict the class membership of the predictor vector x. We wish to build a rule for predicting the class membership of an observation based on p measurements of features X. Traditional methods used are different forms of discriminant rules. The learning data group structure is sometimes not homogeneous. The clusters in the data have often elliptic shape. So, it is a reasonable approach to model the data by mixtures of elliptically symmetric densities. The multivariate normal distribution is extensively used in this area, because of its computational convenience. In practice, the Gaussian assumptions are rarely satisfied. In that case the use of logistic regression is a common practice. This paper focuses on classification when population distributions are not mixtures of Gaussians.

This paper is organized as follows. The formulation of Penalized Multinomial Mixture Logit Model is given in Sect. 2. Results of some simulation studies are presented in Sect. 3. The classification results based on practical life data sets are presented in Sect. 4. Section 5 summarizes our findings of the study and presents some issues for future work.

2 Multinomial logit model

Multinomial logit models are used to model relationships between a polytomous response variable and a set of regressor variables. The term multinomial logit is often used in the econometrics literature to refer to the conditional logit model of McFadden (1974). The term multinomial logit refers to a model that differs slightly from conditional logit model. Theil (1969) in choice of transportation modes and Schmidt and Strauss (1975) in occupational choice of individuals are early applications of the multinomial logit model in the econometrics literature. Schmidt and Strauss (1975) analyzed occupational attainment using the multinomial logit model. The main difference between McFadden (1974)'s conditional logit model and the multinomial logit model is that the multinomial logit model makes the choice probabilities depend on the characteristics of the individuals only, whereas the conditional logit model considers the effects of choice attributes on choice probabilities as well. A detailed explanation is provided in McCullagh and Nelder (1989) and Agresti (1990). McCulloch and Rossi (1994) developed new methods to provide finite sample likelihood based analysis of multinomial probit model.

2.1 Mixtures of multinomial logit model

In a situation when the Gaussian assumptions are not valid and the group structure of the learning data is also heterogeneous, the mixture approach to discriminant analysis can be extended to a case of multinomial logit models. The multinomial logit models are an extension of binary logistic regression models, with a multi-category response matrix instead of binary response in the logistic regression. The idea behind the multinomial logit model is a logistic regression of the categorical response matrix Y on the feature vectors matrix X. In this case one category or group is taken as comparison category. When the classes are also heterogeneous, the indicator response matrix $Y_{n\times J}$ (where each $y_i \sim Mult_J(1,\pi)$ and $\pi = (\pi_1,\ldots,\pi_J)$) in the usual multinomial logit model is replaced with the blurred response matrix $Z_{n\times R}$. Each class j is a mixture of R_j subclasses and $R = \sum_{j=1}^{J} R_j$ is the total number of subclasses. The Z matrix is a mixture analogue of the indicator response matrix except that observations can belong to several subclasses with associated probabilities. If $g_i = j$, i.e., ith observation from the training data falls in the class j, then fill the jth block of R_j entries in the ith row of Z with the values $\hat{P}(C_{jr}|x_i, j)$ and the rest with $0's$, where $\hat{P}(C_{jr}|x_i, j)$ is the cluster probability of an observation i belonging to sub-class r of class j. The random vector $z = (z_1, z_2, \ldots, z_R)$ is assumed to follow a mixture of multinomial distribution. For the fitting of the multinomial logit model we need the subclass membership matrix $Z_{n\times R}$, but only $Y_{n\times J}$ is known. EM algorithm of Dempster et al. (1977) is used for this purpose, the E step being the expectation step of sub-group membership estimation, while the M step is the multivariate iterative re-weighted least squares estimation for the unknown parameters; see, for details, Wilhelm et al. (1998).

2.2 Estimation: EM algorithm

E-Step: The E-step of the EM algorithm gives the estimate of the components of z_i, given the observed x_i and the current fitted parameters at the kth iteration. In the complete data, we need to estimate the sub-class membership z_{ijk} of each object. These z_{ijk} are the current conditional probabilities that x_i belongs to each of the R subclasses. Therefore, the E-step is accomplished by replacing z_{ijk} by its conditional expectation given the observed value x_i and the parameters at the kth iteration, The cluster probabilities are updated as:

$$\begin{aligned} z_{ijk} &= \text{Prob}(z \in \text{kth subclass of } \; j|x_i, j) \\ &= \hat{P}(C_{jk}|x_i, j) \\ &= \frac{P_{ijk}}{\sum_{r=1}^{R_j} P_{ijr}}, \end{aligned} \tag{1}$$

where P_{ijk} are the multinomial probabilities.

M-step: Estimate $\beta's$ by multivariate iterative re-weighted least squares. For the ith observation x_i, the design matrix x_i is a block diagonal of dimension $(R-1) \times p(R-1)$:

$$x_i = \begin{pmatrix} x_{1i}^t \\ \vdots \\ x_{ci}^t \\ \vdots \\ x_{R-1,i}^t \end{pmatrix},$$

where each row x_{ci} consists of the cth block of entries $x_o, \ldots, x_p$, while $0's$ elsewhere.
The link function is defined as:

$$\eta_i = x_i \beta = \begin{pmatrix} \eta_{1i} \\ \vdots \\ \eta_{ci} \\ \vdots \\ \eta_{R-1,i} \end{pmatrix},$$

where each entry η_{ci} models the log odds of subclass c with reference to baseline sub-class R, and the vector β is,

$$\beta = \begin{pmatrix} \underline{\beta_1} \\ \vdots \\ \underline{\beta_c} \\ \vdots \\ \underline{\beta_{R-1}} \end{pmatrix},$$

where each $\underline{\beta_c}$ is a vector of $p + 1$ unknown parameters for each of the $R - 1$ subclasses. The multinomial probabilities constitute the inverse logit link function,

$$h(\eta_i) = p_i = \begin{pmatrix} \frac{e^{(\eta_{1i})}}{1+\sum_{c=1}^{R-1} e^{(\eta_{ci})}} \\ \vdots \\ \frac{e^{(\eta_{R-1,i})}}{1+\sum_{c=1}^{R-1} e^{(\eta_{ci})}} \end{pmatrix},$$

while $1 - p_i = \frac{1}{1+\sum_{c=1}^{R-1} e^{(\eta_{ci})}}$. In this case all the subclass probabilities p_i satisfy a multinomial distribution. The log-likelihood function for the mixture of multinomial logit model in this setting is,

$$l(\beta) = \sum_{i=1}^{n} \sum_{c=1}^{R-1} z_{ic} \ln P_{ic} + \sum_{i=1}^{n} \left[\left(1 - \sum_{c=1}^{R-1} z_{ic} \right) \ln \left(1 - \sum_{c=1}^{R-1} P_{ic} \right) \right] \qquad (2)$$

The score function $\mu(\hat{\beta_k})$ is

$$\mu(\hat{\beta_k}) = \sum_{i=1}^{n} \acute{x_i} \acute{q_i} (U_i - D_i),$$

where U_i is a vector of ratios of z_{ic}/p_{ic} for the first $c = 1, \dots, R-1$ subclasses, D_i is the same ratio for the reference subclass R and q_i is the vector of $\partial h(\eta_i)/\partial(\eta_i)$.
Using Fisher's scoring method, the updated estimates at $k+1st$ iteration are:

$$\hat{\beta}_{k+1} = \hat{\beta}_k + \nu^{-1}(\beta_k)\mu(\hat{\beta}_k),$$

where $\nu(\beta_k)$ is the expected Fisher Information matrix. The parameters of the multinomial logit model are estimated as:

$$\hat{\beta}_{k+1} = \left(\sum_{i=1}^{n} \acute{x}_i w_i x_i\right)^{-1} \sum_{i=1}^{n} \acute{x}_i w_i z_i, \tag{3}$$

with $w_i = D_{p_i} - p_i {p_i}^t$ and D_{p_i} is the diagonal matrix of multinomial probabilities p_i on the main diagonal. Thus the estimates are obtained by a 'multivariate iterative reweighted least squares' *MIRLS* of a working response variable Z on X, where $z_i = x_i\beta_k + z_i^*$ and $z_i^* = U_i - D_i\underline{1}$. The weight matrix W and the adjusted response matrix Z are updated at each iteration, based on current estimates of multinomial probabilities p_i.

2.2.1 Initial values

For starting this EM algorithm, initial estimates are required for the cluster-membership matrix Z, and the parameter vector β. The most obvious way to obtain suitable initial estimates for the parameters when the groups follow a mixture structure is to apply some form of cluster analysis to the data; see, for example, Everitt and Hand (1981). K-Means clustering is used to estimate the cluster probabilities $\hat{P}(C_{jr}|x_i, j)$ for R_j subclasses of each class j, while the estimates of the parameter vector β for each subclass are chosen arbitrarily as $[1, 0, \dots, 0]$.

In the E-step, using the current estimates of the $\beta's$, compute the $\eta_i's$, and hence the multinomial probabilities p_i's. The updates depend on initial values of parameters, and different clustering criterion can produce different initial estimates. So the proposed model estimates are sensitive to the initial clustering method used. The same has been reported in literature; see Hastie and Tibshirani (1996). However, K-Means clustering is used with fixed number of clusters because it does not require prior computation of a proximity matrix of the distance/similarity of every case with every other case. The initial Z matrix is updated using cluster probabilities from K-Means. The M step estimates the parameter vector β, using the current estimates of the weights w_i, working response variable z_i. The EM algorithm is iterated until convergence.

2.2.2 Posterior probabilities

After estimating the parameters for multinomial logit model for the classes, the posterior probability that an unknown observation i belongs to the class j can be estimated using Bayes' theorem. For the general $j = 1, \dots, K$ group classification case, the logistic posterior probabilities for an observation x_i belonging to group j is:

$$P(G = j|x_i) = \frac{e^{x_i'\beta_j}}{1 + \sum_{j=1}^{K-1} e^{x_i'\beta_j}},$$

However, when each class j is a mixture of R_j subclasses, the odds are a mixture of R_j subclasses odds with mixing proportions $\pi_j = (\pi_{j1}, \pi_{j2}, \dots, \pi_{jR_j})$. The posterior probabilities can be generalized as:

$$P(G = j|x_i) = \frac{\sum_{r=1}^{R_j} \pi_{jr} e^{x_i'\beta_{jr}}}{1 + \sum_{c=1}^{R-1} \pi_c e^{x_i'\beta_c}},$$

where π_c are the mixing proportions for the $c = R - 1$ subgroups such that for each group j, $\sum_r \pi_{jr} = 1$, and the β's are estimated from the M-step of the multinomial logit model. For the bivariate binary classification problem, where each group is a mixture of R_j subgroups, the posterior probability can be computed for an observation belonging to group 1 as,

$$P(G = 1|x_i) = \frac{\pi_{11} e^{x_i'\beta_{11}} + \pi_{12} e^{x_i'\beta_{12}} + \cdots + \pi_{1R_1} e^{x_i'\beta_{1R_1}}}{1 + \sum_{c=1}^{R-1} \pi_c e^{x_i'\beta_c}},$$

where $x_i' = (1, x_{i1}, x_{i2})'$. This proposed method of logistic regression in mixture models is different from Mixtures of Experts Models, where the conditional distribution of the responses (given covariates) is considered as a mixture of Generalized Linear Models; see, for example, Peng et al. (1996).

2.3 Computational issues

The mixture model problem of two groups, where each group was a mixture of two subgroups was set up using multinomial logit models. The approach was to regard x as fixed and take the unobserved subclass membership vector z to have multinomial distribution. But a problem was encountered in the estimation. Because the subclasses were separated by K-means clustering that was used to get initial estimates of subgroup membership $\hat{P}(C_{jr}|x_i, j)$, so the cluster probability was either 0 or 1, resulting in an infinite log-likelihood for the multinomial logit model and infinite parameter estimates. The estimation is sensitive to initial values of estimates, however, if we are somehow able to take initial $\hat{P}(C_{jr}|x_i, j)$ other than (0, 1), we can solve this problem. The same problem has been reported in literature; see, for example, Ripley (1996). The maximum likelihood has an infinite component if some groups can be completely separable on a linear projection or quasi-complete separable from others. In this case, the multinomial logit model has infinite slope. In logistic regression it has been recognized that with small to medium-sized data sets, situations may arise where, although the likelihood converges, at least one parameter estimate is infinite; Albert and Anderson (1984). These situations occur if the responses and non-responses can be perfectly separated by a single risk factor or by a non-trivial linear combination of risk factors. Therefore Albert and Anderson (1984) denoted such situations by 'separation'. The phenomenon is also known as 'monotone likelihood'. In general, one does not assume infinite parameter values in underlying populations. The problem of separation is rather one of non-existence of the maximum likelihood estimate under special conditions in a sample. An infinite estimate can also be regarded as extremely inaccurate.

2.4 Penalized multinomial logit model

To overcome the problem caused by the complete separation between the subgroups, the concept of penalized multinomial mixture logit model was introduced with a penalty inducted in to bring down the infinite component of the maximum likelihood estimators. Good and Gaskins (1971) were the first to introduce the idea of roughness penalty estimation in density estimation. Hoerl and Kennard (1970a) presented the idea of ridge regression which is a simple form of penalized regression, to cope with multicollinearity of regressors in case of linear regression using quadratic penalty. Other penalties lead to lasso of Tibshirani (1996) or to bridge regression of Frank and Friedman (1993). The penalized or shrinkage methods have a long history in many fields since 1970 such as linear discriminant analysis (Friedman 1989), logistic regression (Schaefer et al. 1984). Heinze and Schemper (2002) provided the solution to the problem of separation using modified score function of Firth (1993). The penalized log-likelihood in the case of proposed model is:

$$l^*(\beta) = l(\beta) - 0.5\lambda\beta'\Omega\beta,$$

where $\lambda > 0$ is the ridge parameter that controls the size of the coefficients (increasing λ value decreases their size) and Ω is a $p(R-1) \times p(R-1)$ block diagonal penalty matrix with entries $\Omega_c = Var(\underline{X})$ in each of $R-1$ blocks, while the entries corresponding to the parameter $\beta_\circ$ are set as zero. This sets the regression coefficients to penalization, not the offset $\beta_\circ$. In this setting the ridge coefficient λ is a perturbation of the diagonal entries of $X'WX$ to encourage non-singularity. Then the penalized multinomial mixture logit model parameter estimation is performed, replacing Eq. (3) with Eq. (4). This results in a biased estimate $\hat{\beta}_{k+1}$.

$$\hat{\beta}_{k+1} = \left(\sum_{i=1}^{n} \acute{x}_i w_i x_i + \lambda\Omega\right)^{-1} \sum_{i=1}^{n} \acute{x}_i w_i z_i, \tag{4}$$

Penalization makes the norm of the coefficient matrix smaller, helping avoid overfitting problem. The general idea behind penalization is to avoid arbitrary coefficients estimates. The choice of ridge parameter λ is crucial and has been under discussion in many contexts in literature, the most widely used approach is the cross validation *CV*. Criterion for the performance of *CV* are either misclassification rate or the strength of prediction. In our proposed model as classification of entities is the main concern, we tried different values of λ and the value that was able to tackle the problem of separable subclasses and thus shrinking the log-likelihood for finite parameter estimation and giving smaller errors was chosen. However, further study can be conducted for the choice of λ in the context of proposed penalized multinomial mixture logit models (pmml).

3 Simulations

This section describes the results of different sets of simulations performed to study the performance of our model as compared to the classical methods of discrimination.

3.1 Multivariate normal data

This simulation was carried out using a mixture of two bivariate Gaussian components. Each of the two groups consisted of two multivariate Gaussian subgroups. The sub-group means were: $\mu_{11} = [1.5\ 1.5]'$, $\mu_{12} = [-1.5\ 1.5]'$, $\mu_{21} = [1.5\ -1.5]'$, and $\mu_{22} = [-1.5\ -1.5]'$. The training sample had 80 observations with equal priors for the two groups, while the test sample size was 40. The common covariance was the identity matrix. The errors of misclassification for the training and the test data are presented in Table 1. In this table, the first row represents the errors of misclassification using penalized multinomial mixture logit models (pmml), while the second row represents errors, using mixture discriminant analysis (mda) approach of Hastie and Tibshirani (1996). The last two rows also report the simulation results from linear and quadratic discriminant analysis (lda) and (qda), respectively. The values reported in Table 1 are averages of misclassification errors over 30 simulations, with the standard error of the average in parentheses. The errors are much higher in both these models, as the distributional assumptions are not satisfied. As the data were generated from mixture of Gaussian distributions, so the approach of Hastie and Tibshirani (1996) worked much better as compared to our proposed model. The mixture of penalized logit models was tried with different values of the shrinkage parameter λ and the value of $\lambda = 0.00025$ was chosen for this simulation. However the existence of solution for penalized multinomial mixture logit model depends on the value of λ that is able to reduce the effect of separation of subclasses due to K-Means clustering in the initial stage. However, each simulation would produce different configuration of the n sample points. So repetition of the simulations, with a fixed value of λ to penalize the regression coefficients was not possible a large number of times.

Table 1 Errors of misclassification for multivarite normal data

Methods	Training	Test
pmml	0.0594 (0.0324)	0.0913 (0.0383)
mda	0.0038 (0.0100)	0.0088 (0.0147)
lda	0.2350 (0.0483)	0.2662 (0.0545)
qda	0.2650 (0.0545)	0.2662 (0.0409)

Table 2 Errors of misclassification for Example 1 data set

Methods	Training	Test
pmml	0.4050 (0.0453)	0.4492 (0.0641)
mda	0.4446 (0.0381)	0.4967 (0.0966)
lda	0.4609 (0.0457)	0.4933 (0.1032)
qda	0.3862 (0.0559)	0.4525 (0.0893)

It is clear from Table 1 that as data were generated from mixture of Gaussian distributions, so the standard *mda* approach performed far much better than the proposed penalized multinomial mixture logit model (*pmml*).

3.2 Multivariate logistic data

This section presents different simulation examples for mixture of logistic data generated with certain degree of distinctness and compares the performance of the penalized multinomial mixture logit model (pmml) with other classical methods.

Example 1 Here we sampled mixtures of bivariate logistic data from location set $[-1, -1] \times [1, 1]$ and scale being fixed as $[2, 1]$.

The test error rates for this simulation are shown in Table 2. The values reported in Table 2 are averages of errors over 30 simulations, with the standard error of the average in parentheses. In this case, both our proposed model and the *qda* work better than *mda* and *lda*, as quadratic or non-linear discriminant function better describe the decision boundaries in this type of data structure as also shown in Fig. 1. The decision boundary produced using our model is very wiggly, so smoothed functions of decision boundary from our model are plotted. Over-specifying the number of subgroups assuming no a prior knowledge of the number of subgroups in the mixture caused estimation problems. The value of $\lambda = 0.016$ was chosen from a grid of penalties between 0 and 1, producing minimum test errors. The test errors for this grid show in general a rising pattern for higher values of λ as shown in Fig. 2.

Example 2 This simulation was carried out using a mixture of two logistic distributions. Each of the two groups was generated as a mixture of two logistic components. The sub-group means were: $\mu_{11} = [-1.5 \quad -1.5]'$, $\mu_{12} = [1.5 \quad -1.5]'$, $\mu_{21} = [-1.5\ 1.5]'$, and $\mu_{22} = [1.5\ 1.5]'$, while the scale was assumed identity. The training sample comprised of 80 observations with equal priors for the two groups, while the test sample size was 40. The performance of our proposed model was slightly poor as compared to other methods. In this situation, none of the non-linear methods significantly outperform the lda in terms of test error rates. The classes seem to have linear separation shown in Fig. 3. Different values for shrinkage parameter were tried, but $\lambda = 0.0008$ resulted in the test errors reported in Table 3 using the proposed model. The values reported in Table 3 are averages of errors over 30 simulations, with the standard error of the average in parentheses. The test errors plotted as a function of λ are shown in Fig. 4. We also tried to over-specify the number of subgroups, but faced estimation problems.

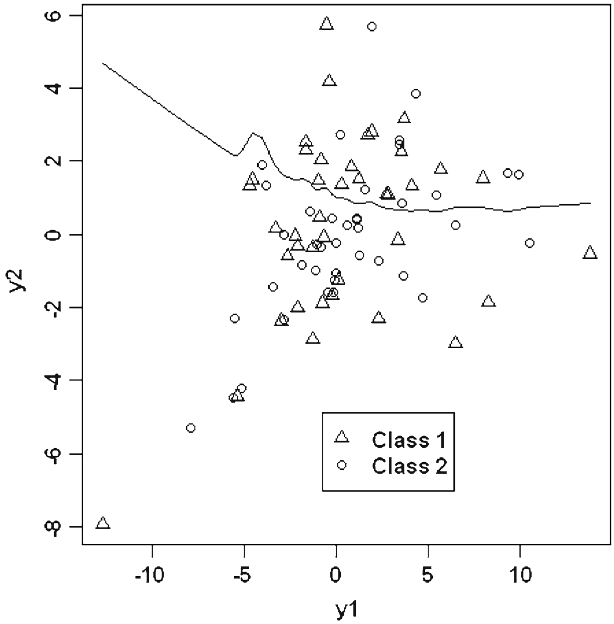

Fig. 1 Data structure for mixtures of bivariate logistic groups classification problem of Example 1. Here the smoothed function of the decision boundary produced by *pmml* is shown

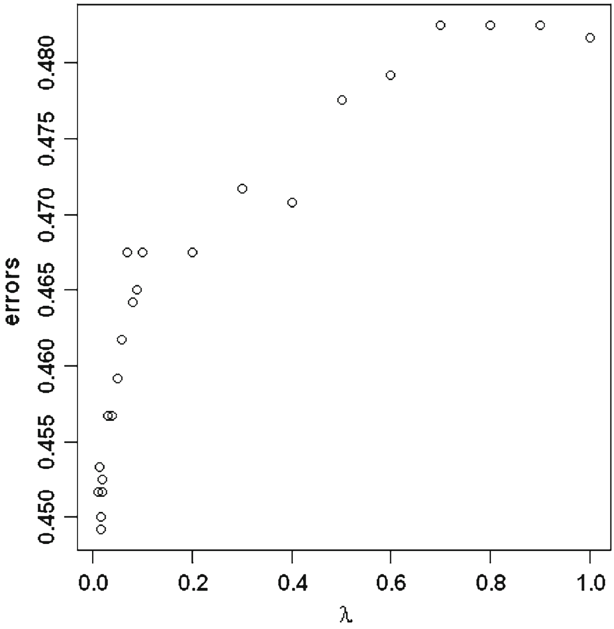

Fig. 2 Choice of λ for the classification problem of Example 1

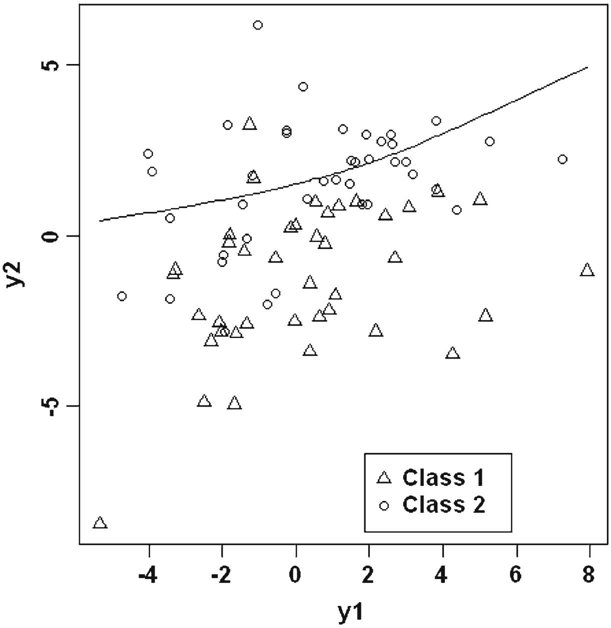

Fig. 3 Example 2 data structure along with decision boundary of *pmml*

Table 3 Example 2: errors of misclassification

Methods	Training	Test
pmml	0.1758 (0.0491)	0.1950 (0.0617)
mda	0.1733 (0.0605)	0.1742 (0.0464)
lda	0.1667 (0.0480)	0.1742 (0.0596)
qda	0.1675 (0.0499)	0.1725 (0.0634)

Example 3 In this simulation, the logistic components means were: $\mu_{11} = [2.5\ 2.5]'$, $\mu_{12} = [0\ 0]'$, $\mu_{21} = [1\ 1]'$, $\mu_{22} = [0.5\ 0.5]'$, while the scale vector was fixed as (0.5, 0.25). The first class almost completely surrounds the second class as shown in Fig. 5. The training and the test sample was of the same structure as in previous simulation. Different values for shrinkage parameter were tried, but $\lambda = 0.004$ resulted in minimum test errors. The results are reported in Table 4, where the values reported are averages of errors over 30 simulations, with the standard error of the average in parentheses. The performance of our model is much better than all the three classical methods producing minimum test errors of 0.2087. Smaller values of λ in this case resulted in smaller test errors, as shown in Fig. 6.

Example 4 In this simulation, the logistic components means were: $\mu_{11} = [0\ 0]'$, $\mu_{12} = [1.5\ 1.5]'$, $\mu_{21} = [0\ 0]'$, $\mu_{22} = [-1.5\ 1.5]'$, while the scale was randomly generated from Uniform (0,1). The training and the test samples were of the same structure as in previous simulation. Different values for shrinkage parameter were tried, but $\lambda = 0.01$ resulted in minimum test errors. Our proposed model *pmml* did not

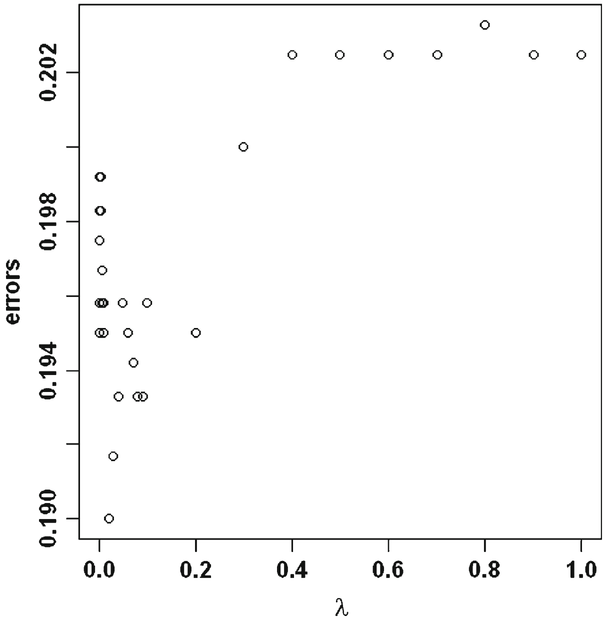

Fig. 4 Choice of λ in Example 2. Here each point represents test errors produced for a roughness penalty parameter value

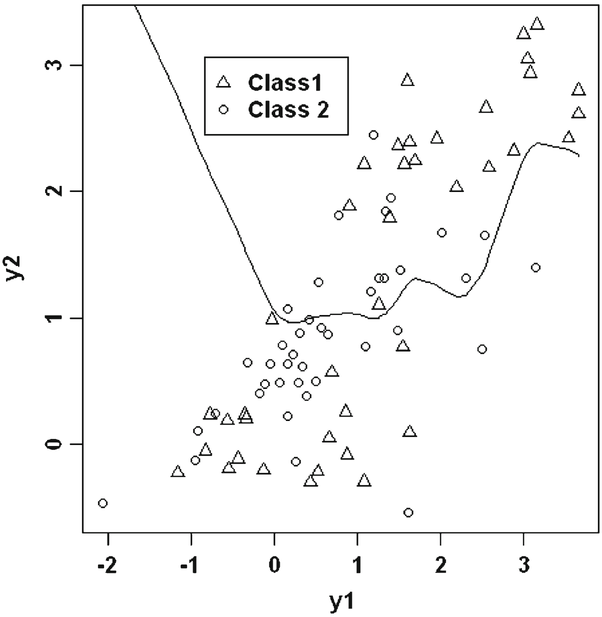

Fig. 5 Example 3 data classification problem: class 1 almost completely surrounds the class 2

Table 4 Example 3: errors of misclassification

Methods	Training	Test
pmml	0.1869 (0.0564)	0.2087 (0.0650)
mda	0.4844 (0.0508)	0.5175 (0.0994)
lda	0.3775 (0.0645)	0.3962 (0.0886)
qda	0.2106 (0.0644)	0.2600 (0.0710)

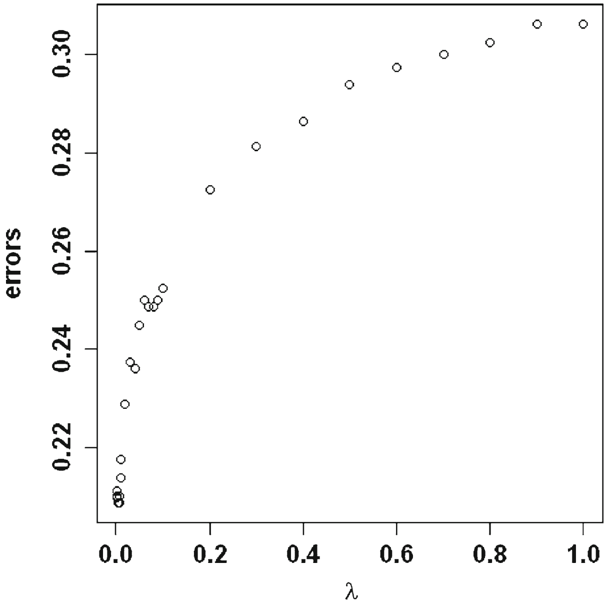

Fig. 6 Choice of λ for simulation Example 3

Table 5 Example 4: errors of misclassification

Methods	Training	Test
pmml	0.2513 (0.0598)	0.2808 (0.1006)
mda	0.2629 (0.0649)	0.2825 (0.0785)
lda	0.2542 (0.0582)	0.2708 (0.0799)
qda	0.2550 (0.0651)	0.2608 (0.0827)

outperform than all the three classical methods as is shown from test errors in Table 5. Figure 7 displays graphically the data structure with decision boundary, while Fig. 8 shows test errors plotted as a function of λ. Figure 8 gives no evidence of any pattern between test errors as a function of λ. So choice of λ in this example was random.

Example 5 (3 Groups) To study the performance of this method in a complex data structure, we generated data for 3 groups. Each of the three groups were generated as a mixture of two logistic components. The sub-group means were: $\mu_{11} = [2.5\ 2.5]'$,

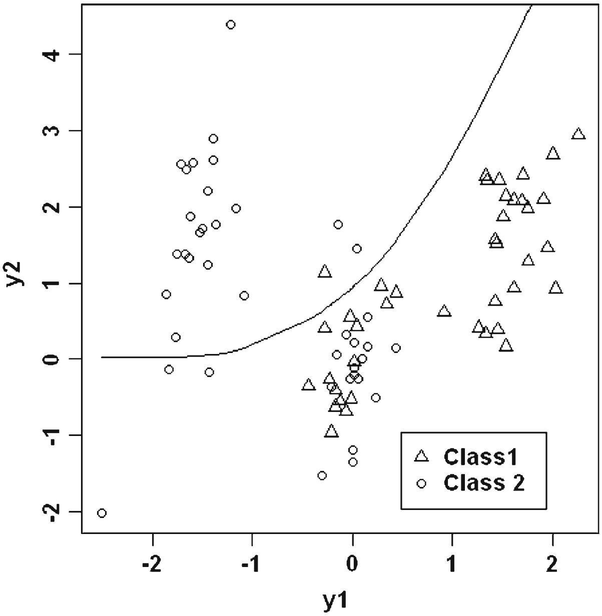

Fig. 7 Example 4 data Classification. There is an overlap between the points from the two classes in the middle region

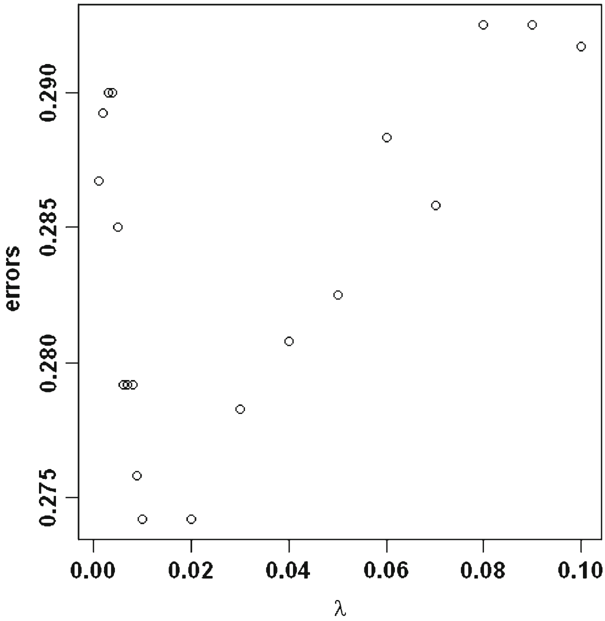

Fig. 8 Choice of λ for Example 4 data classification problem

Table 6 Three groups classification problem: errors of misclassification

Methods	Training	Test
pmml	0.2475 (0.0491)	0.2844 (0.0685)
mda	0.5264 (0.0915)	0.5389 (0.0966)
lda	0.3825 (0.0412)	0.3939 (0.0579)
qda	0.3344 (0.0532)	0.3467 (0.0542)

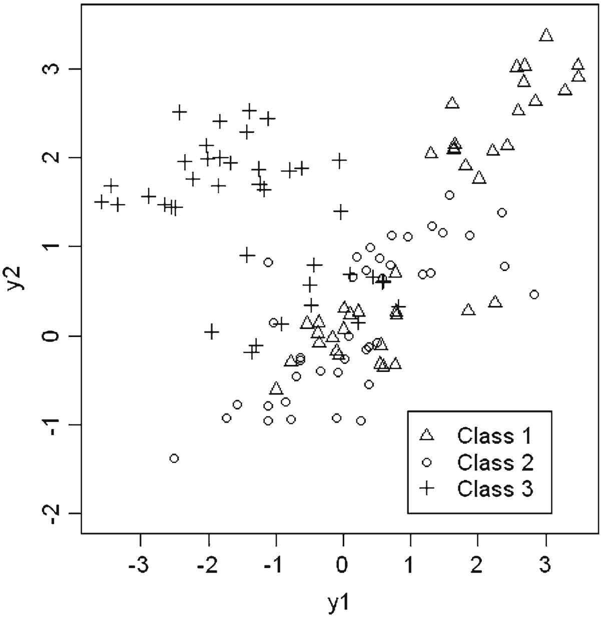

Fig. 9 Data structure for Example 5 with three groups

$\mu_{12} = [0\ 0]'$, $\mu_{21} = [1\ 1]'$, $\mu_{22} = [-0.5\ -0.5]'$, $\mu_{31} = [-2\ 2]'$, $\mu_{32} = [-0.5\ 0.5]'$, while the scale vector was (0.5 0.25). The training sample comprised of 120 observations with equal priors for the three groups, while the test sample size was 60. The proposed model did very well as compared to other methods, as data were generated from mixtures of logistics. The test errors are reported in Table 6, where the values reported are averages of errors over 30 simulations, with the standard error of the average in parentheses. At $\lambda = 0.001$ our model produced minimum test errors of 28% and was the optimal model chosen. The data structure for this example is shown in Fig. 9.

All the simulation examples presented here show a general message about the choice of penalty parameter λ, that smaller values of λ produce smaller test errors, as is also intuitive that penalization produces biased estimates, smaller the value of λ, better will be the classification.

Table 7 Errors of misclassification, t-distributions

Methods	Training	Test
pmml	0.2105 (0.0796)	0.3000 (0.0354)
mda	0.2667 (0.0059)	0.3563 (0.0619)

3.3 Mixtures of multivariate t-distributed data

For discrimination purposes, unstructured data were generated from a mixture of multivariate t-distributions, so there were outlying observations. We wanted to test the performance of our proposed method as compared to standard *mda* approach in the presence of outliers. There were four groups and each group was a mixture of three spherical bivariate normal subgroups, with a standard deviation of 0.25. The means of 12 subclasses were chosen at random (without replacement) from the integers $(1, \dots, 5) \times (1, \dots, 5)$ and the degrees of freedom for each subclass was chosen as 5. Each subclass was comprising of 20 observations, with a total of 240 observations in the training sample and test sample was of size 80; see, for example, Hastie and Tibshirani (1996). The data have extremely disconnected class structure, so we expect relatively higher misclassification errors. The errors of misclassification using the *pmml* and *mda* approaches are presented in Table 7. With a value of $\lambda = 0.000025$, we were able to estimate our parameters, but we were not able to perform a large number of simulations, due to inability to estimate parameters with a fixed value of λ. However, we think that this problem might be tackled with a better programming facility, as well as using cross-validation approach. Table 7 shows that using the proposed *pmml* model more accurate classification results were obtained with a test error of 0.30 as compared to 0.35 using *mda* approach. This was expected, as data were generated from heavier tailed mixtures of multivariate t-distributions, so *mda* which is based on the assumption of multivariate normality breaks down. This also shows that our model is more robust to outliers than *mda* model.

4 Results

The proposed penalized multinomial mixture logit model (pmml) is applied on two different data sets to compare the classification performance of this method with other classical methods of discrimination such as linear discriminant analysis (lda), quadratic discriminant analysis (qda), etc.

4.1 Example: forensic glass data

This example is from forensic testing of glass. The glass data were obtained from the UCI Machine Learning Repository maintained by Murphy and Aha (1995). After a careful examination of the data, we found one of the variable was having repeated values of 0, so we ignored this variable to avoid any unseen problems. We chose 7 predictors defined in terms of their oxide content (i.e. Na, Mg, Al, etc.) while leaving

Table 8 Errors of misclassification, forensic glass test data

Methods	Errors (2-sub)	Errors (3-sub)
pmml	0.3167	0.2500
mda	0.4333	0.4000[a]

[a] $lda = 0.3833$, $qda = 0.4833$

out ID and refractive index. The training data consisted of two groups and 7 predictors. The two groups are window float glass and window non-float glass. The variables measured are weight proportions of different oxides. A sample of 80 observations with equal priors for the two groups was chosen as the training set, while the test data were of size 60. Assuming that the two groups are a mixture of two subgroups each, the data were analyzed using *mda* as well as (pmml) models. The penalty parameter $0 < \lambda < 2$ was tried. For $\lambda = 0.55$, the test errors were least using each group as a mixture of two subgroups; see Table 8. With the assumption of three subgroups, $\lambda = 1.5$ produced much improved classification with errors of 0.25 as compared to the errors of 0.40 by *mda*. The classical methods of discrimination *lda* and *qda* did not perform well for this dataset with the test errors of 0.3833 and 0.4833 respectively. Therefore, the *pmml* model with three subgroups within each group performed exceptionally well as compared with all other classification methods tried.

4.2 Example: leukemia data

A very important use of discrimination methods is their application in image diagnostics and successful treatments afterwards. In practice the disease groups are not homogeneous, as most of the discrimination methods assume. Then it seems natural to think the groups as a mixture of subgroups and the application of models based on mixture of distributions is quite rational. The proposed penalized multinomial mixture logit model was applied to the leukemia data set of Golub et al. (1999) and obtained from http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. This data set consists of gene expressions of 72 patients of two types of leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The ALL class is heterogeneous and consists of 38 cases of type B-cell ALL and 9 cases of T-cell ALL, while the AML class consists of 25 cases. Three preprocessing steps for filtering of genes were applied to the data; see, for details, Dudoit et al. (2002). Furthermore, a large number of genes do not show variability across groups, so their contribution to classification is not significant. The most relevant $p = 40$ genes contributing most to the classification were chosen by the ratio of between-group to within group sums of squares, based on the learning set of 38 cases of Golub et al. (1999). In the learning set of 38, first 27 cases are of 2 types of ALL, while the last 11 cases belong to AML class leukemia. Then *pmml* model was fitted to the training set of 38 patients and parameters estimated assuming that the both ALL and AML classes were a mixture of two subgroups each, so a total of $R = 4$ subgroups. Different values of λ were tested for estimation purposes and $\lambda = 0.0025$ was chosen producing optimum classification results for the test data of 34 observations. The comparison was made with Diagonal Linear Discriminant Analysis (dlda) and Diagonal Quadratic Discriminant Analysis (dqda). The results are

Table 9 Comparison of classification for leukemia data set

Methods	Test error
*pmml*4	0.0588
*pmml*3	0.0294
dlda	0.0294
dqda	0.4412

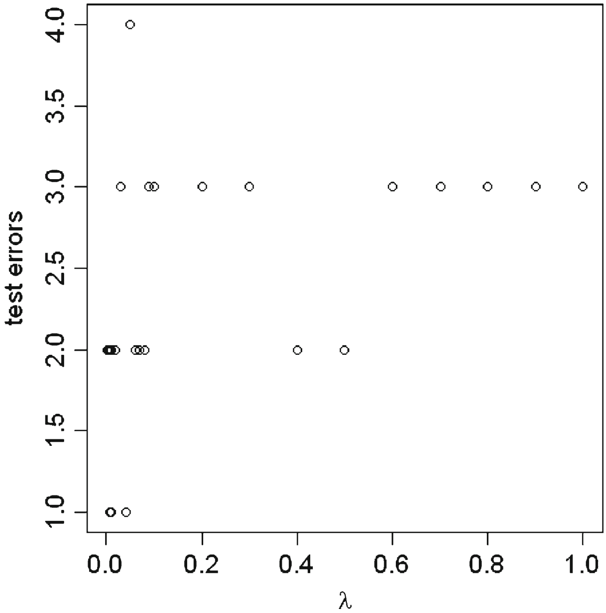

Fig. 10 Choice of λ for leukemia data analysis

presented in Table 9. It is clear from Table 9 that *dlda* performed exceptionally well, but the performance of penalized multinomial mixture logit model (*pmml4*) was also encouraging. Here *pmml4* denotes a penalized multinomial mixture logit model with four subgroups. Figure 10 graphically displays the choice of λ in this example. It is evident from Fig. 10 that test errors show no fixed trend in relation to the value of the shrinkage parameter λ.

Next the penalized multinomial mixture logit model was fitted to the training data of 38 observations, but now assuming that just the ALL class is heterogeneous, not the AML class, so a mixture of $R = 3$ subclasses. The value of λ that produced accurate classification was 0.3. The results were comparable to *dlda*; see Table 9. With this assumption, using *pmml3*, the test data error rate was almost 3%. Here *pmml3* denotes a penalized multinomial mixture logit model with 3 subgroups. The only observation that was misclassified by a very small margin was actually on the border line of the group ALL and AML. So *pmml3* was equally efficient model to *dlda* in this case. For

Table 10 Out-of-sample errors for leukemia data set

Methods	Median test error	Standard deviation
$pmml4$	1	0.6645
$dlda$	0	0.4726
$dqda$	1	1.0222

this data set, observation number 66 from the test set was misclassified by *dlda* and observation 67 misclassified from the test set using *pmml3*.

Then a re-randomization study was performed, i.e., an out of sample analysis on 100 random subdivision of the data set into a learning set of 48 observations and a test set of 24 observations. For each subdivision, the 48 learning set observations were chosen giving proportional weights to ALL and AML classes, to overcome the problems of estimation of mixtures of subgroups. Again a mixture of 4 subgroups penalized multinomial logit model (pmml4) was fitted to each learning set and then test errors computed for each sub-division. The value of the ridge parameter $\lambda = 0.0025$, was chosen after a number of trials testing different values. Again three different discriminant rules were tested for their classification performance, i.e., *pmml*, *dlda*, *dqda*. The results in Table 10 are the summary of classification errors over 100 random sub-divisions of the data set.

From Table 10, it is clear that the median error rate for *dlda* is minimum, as also reported by Dudoit et al. (2002), but the performance of our penalized model is also quite encouraging, the observation number 66 was misclassified by all the methods, whenever it was in the test sample, but *pmml4* and *dqda* also misclassified observation 67. However, in the re-randomization study of the leukemia data, *pmml4* performed better than the *dqda*, while *dlda* being the best; see, Table 10. The performance of *pmml* with three subgroups on the original test set of Leukemia data of Golub et al. (1999) was similar to that of *dlda*, this method is recommended for classification problems involving heterogeneous groups.

4.3 Example: breast cancer data

The breast cancer database was obtained from machine learning databases of Wolberg and Mangasarian (1990). There are two groups and 33 variables. Each of the chosen group has a total of 47 cases. Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Wolberg and Mangasarian (1990) since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The first 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Relevant features were selected using an exhaustive search in the space of 1–4 features and 1–3 separating planes; see, for detailed information about the attribute information, Wolberg and Mangasarian (1990). The results of the analysis of these data are recorded in Table 11. The first column shows the method of discrimination applied. The $pmml2$ is the proposed penalized multinomial mixture logit model with two subgroups, while lda and qda

Table 11 Errors for breast cancer data set

Methods	Error
*pmml*2	0
lda	0.0745
qda	0

are the traditional methods of classification. The values are the error probabilities using resubstitution method. From Table 11, it is clear that performance of *pmml2* was as encouraging as that of *qda* compared to *lda*, which misclassified 7 cases. Further, with the application of classical methods certain rank issues also arise, as these methods set an upper limit on the number of variables available as compared to the number of cases, but our model does not have such limitation.

5 Conclusion

This paper applies the concept of penalized logistic regression for multinomial mixture logit models, to several data sets. We have shown that the use of penalized multinomial mixture logit models help in the improvement of classification performance, when the groups are not homogeneous. The proposed model performed better when the data were generated from mixtures of logistics or t-distributions, as well as in the Forensic Glass data problem and two cancer data sets classification. Though, in our random sub-division study on leukemia data set, *dlda* performed a little better than the proposed model, but the *dlda* has a drawback of ignoring the correlations between genes, which are also important biologically; see, for example, Dudoit et al. (2002). Furthermore, we addressed only the problem of classification, using one method for selecting desirable genes, but it can be tried using different methods of gene selection. Further work needs to be done regarding the choice of the penalty parameter λ, which plays crucial role in the working of our proposed model. Though, we fixed the value of λ based on the classification results of Golub et al. (1999) test set, but it needs to be adjusted for each random sub-division, because the configuration of sample points in the learning set changes for each sub-division. Future work in this area needs to address the issue of smart choice of penalty parameter λ, which helps in overcoming the problem of separate subclasses resulting in infinite parameter space.

Acknowledgments The breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Wolberg and Mangasarian (1990).

References

Agresti A (1990) Categorical data analysis. John Wiley & Sons, Inc, New York

Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71:1–10

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38

Dudoit S, Fridlyand J, Speed T (2002) Comparison of discrimination methods for the classification of tumors using GENE expression data. J Am Stat Assoc 97(457):77–87

Everitt BS, Hand DJ (1981) Finite mixture distributions. Chapman & Hall, London

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27–38

Frank IE, Friedman JH (1993) A Statistical view of some chemometric regression tools. Technometrics 35:109–148

Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84:165–175

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Good IJ, Gaskins RA (1971) Nonparametric roughness penalties for probability densities. Biometrika 58:255–277

Hastie T, Tibshirani R (1996) Discriminant analysis by Gaussian mixtures. J R Stat Soc Ser B 58(1):155–176

Heinze G, Schemper M (2002) A solution to the problem of separation in logistic regression. Stat Med 21:2409–2419

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

McCulloch R, Rossi PE (1994) An exact likelihood analysis of the multinomial probit model. J Econom 64:207–240

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka PFrontiers in econometrics. Academic Press, New York pp 105–142

Murphy PM, Aha DW (1995) UCI repository of machine learning databases dept of information and computer science, University of California, Irvine, California. http://www.ics.uci.edu/~mlearn/MLRepository.html

Peng F, Jacobs RA, Tanner MA (1996) Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. J Am Stat Assoc 91(435):953–960

Ripley BD (1996) Pattern recognition and neural networks. University Press, Cambridge

Schaefer R, Roi L, Wolfe R (1984) A ridge Logistic estimator. Commun Stat Theory Methods 13(1):99–113

Schmidt PJ, Strauss RP (1975) The prediction of occupation using multiple logit models. Int Econ Rev 16:471–486

Theil H (1969) A multinomial extension of the linear logit model. Int Econ Rev 10:251–259

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288

Wilhelm MS, Carter EM, Hubert JJ (1998) Multivariate iterative re-weighted least squares, with applications to dose–response data. Environmetrics 9:303–315

Wolberg WH, Mangasarian OL (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci USA 87:9193–9196. [ftp://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/]