Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Multivariate Analysis 93 (2005) 102–111
Journal of Multivariate Analysis
http://www.elsevier.com/locate/jmva

High breakdown mixture discriminant analysis

Shaheena Bashir* and E.M. Carter

Department of Mathematics and Statistics, University of Guelph, Guelph, Ont., Canada N1G 2W1

Received 6 August 2002

Abstract

Robust S-estimation is proposed for multivariate Gaussian mixture models generalizing the work of Hastie and Tibshirani (J. Roy. Statist. Soc. Ser. B 58 (1996) 155). In the case of Gaussian Mixture models, the unknown location and scale parameters are estimated by the EM algorithm. In the presence of outliers, the maximum likelihood estimators of the unknown parameters are affected, resulting in the misclassification of the observations. The robust S-estimators of the unknown parameters replace the non-robust estimators from M-step of the EM algorithm. 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 robust S-estimators as compared to the standard maximum likelihood estimators.

AMS 2000 subject classifications: 62F35; 62H30

Keywords: Mixture models; EM algorithm; S-Estimators; Breakdown point

1. Introduction

Fisher's linear discriminant analysis is a time-honored rule for the classification of objects with p feature variables, into different source populations. Much work has been carried out and developments made in the area of discriminant analysis; see, for example [10]. Therefore, there are many allocation rules available, depending on the type of situation faced by the experimenter.

*Corresponding author.
E-mail address: sbashir@uoguelph.ca (S. Bashir).

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2003.12.003

1.1. Mixture discriminant analysis (mda)

In the mixture approach to discrimination, it is assumed that we have n_j training observations from population j, $j = 1, \ldots, J$. Each of the population j is divided into R_j artificial subclasses denoted by C_{jr}, $r = 1, \ldots, R_j$ and define $R = \sum_j R_j$, where $n = \sum_j n_j$. According to this clustered approach, an observation from the rth subclass of the jth population has a multivariate normal distribution with its own mean vector μ_{jr} and common covariance matrix Σ. The prior probability for population j is π_j. Let π_{jr} be the mixing probability for the rth subclass in the jth population, such that $\sum_r \pi_{jr} = 1$. The mixing proportions π_{jr} are unknown model parameters, while priors are known or easily estimated from the data. The mixture density for the population j is

$$m_j(x) = P(X = x|G = j) = |2\pi\Sigma|^{-\frac{1}{2}} \sum_{r=1}^{R_j} \pi_{jr} \exp[-D(x, \mu_{jr})/2], \tag{1}$$

where $D(x; \mu_{jr})$ is the Mahalanobis distance between x and μ_{jr} with respect to Σ. EM algorithm is used for the estimation of the parameters; see, for example [2]. The E-step gives an estimate of cluster probability $\hat{P}(C_{jr}|x_{ij}, j)$, while the estimates of the location vectors for different clusters and the common covariance matrix are obtained at the M-step; see, for details [4]. The EM algorithm requires a choice of number of subgroups R_j, starting values for μ_{jr}, Σ and the cluster probabilities $P(C_{jr}|x_{ij}, j)$. A lot of work has been carried out for the selection of the number of clusters; see, for example, [14]. The number of subgroups are chosen to minimize the errors of classification for the test data; see [4]. There are different strategies for getting these values. For each class j, after choosing a fixed number of clusters, the K-means clustering algorithm would be used to estimate a set of R_j subclass centroids $\tilde{\mu}_{jr}$. The initial estimate of Σ would be obtained by pooling together within cluster covariance matrices.

The posterior probability for the jth class is estimated by

$$P(G = j|X = x) \sim \pi_j \,\mathrm{Prob}(x\,|\,j) \sim \pi_j \sum_{r=1}^{R_j} \pi_{jr} \exp[-D(x, \mu_{jr})/2]. \tag{2}$$

An observation is classified into the class j with maximum posterior probability. The classification rules depend on the unknown parameters, which are to be estimated from the training data. In the presence of a number of outlying observations in the training data, the estimates of the unknown parameters can be unstable due to the undue influence of these atypical observations. High breakdown estimation is a procedure designed to remove this cause of concern, by producing estimators that are robust to serious distortion by outliers, eliminating the influence of such atypical observations. However, it is an important fact that in discriminant analysis, not only are the outliers a concern but also inliers. In the K-means clustering, the outliers for one group might be the inliers for others affecting the classification performance, while in case of mixtures of distributions, this situation may be even worst.

1.2. High breakdown estimation

Robust methods for the estimation of multivariate location vector and covariance matrix, have been under development for many years. The M-estimators of $\mu_{p\times 1}$ and $\Sigma_{p\times p}$ were defined in [9]. The breakdown point of M-estimators is at most $1/(p+1)$, due to the increasing sensitivity of covariance M-estimators to outliers contained in lower-dimensional hyperplanes as p gets larger; see, for example, [7,9]. The idea of model fit diagnostics was given in [5], but this method is not reliable, as sever multivariate outliers may be left undetected. The minimum volume ellipsoid (MVE) and minimum covariance determinant estimators (MCD) were defined to guard against multiple outliers in higher dimensions; see, for example, [12].

The S-estimators of multivariate location vector and scatter matrix were defined in [13] having higher breakdown point as compared to the M-estimators in higher dimensions. The idea of robust M-estimation was extended to the mixture models; see, for example, [11], whereby observations assessed as atypical of a component or the mixture itself are given reduced weight in the computation of the estimate of the parameter ϕ of the mixture. It was shown in [8] that S-estimators are in the class of M-estimators with standardizing constraints. The robust S-estimators of multivariate location vectors and common dispersion matrix, were used; see [6] to discriminate between the two populations. The influence function for the estimators of the parameters of the discriminant function and for the associated classification error was worked out, [1]. The multiple outliers are hard to identify in multivariate data clouds, creating a situation of masking and swamping. The conventional maximum likelihood estimators are affected by the presence of outliers, and so break down. These non-robust estimators influence the discriminant function, leading to the poor classification.

In Section 2, we discuss multivariate normal mixture models having outliers and our proposal to use the high breakdown point S-estimators in the M-step of the EM algorithm. A part of our simulation results are given in Section 3. A comparison of the efficiencies of the two methods via simulations is also presented in Section 3.

2. S-estimators in mixture model

In the mixture models, the parameters estimates obtained are the maximum likelihood estimators of the location vectors and the common covariance matrix. In the presence of outliers, these estimators are non-robust. Our proposal is to replace the maximization step, the M-step of the EM algorithm with a robust S-estimation step.

Definition 1. In the case of mixture model, considering the pooled sample, the S-estimators are the vector $\tilde{\mu}_{jr}$ and the positive-definite matrix $\tilde{\Sigma}$ that minimize $|\Sigma|$ subject to

$$n^{-1}\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)\rho[\{(x_{ij}-\mu_{jr})'\Sigma^{-1}(x_{ij}-\mu_{jr})\}^{1/2}] = K_p \tag{3}$$

among all $(\tilde{\mu}_{jr},\tilde{\Sigma})\in\theta$, where θ is the parameter space.

For high breakdown point, choose a tuning constant $c_\circ$ such that $K_p/\rho(c_\circ) = r$, where r is the limiting value of the finite sample breakdown point and $0 < K_p < \rho(c_\circ)$. This is an extra condition on the function ρ. Define $K_p = E_\Phi[\rho||Z||]$, where Z follows a standard multivariate normal distribution. Tukey's biweight function is the most popular choice for the ρ function. It gives redescending ψ function. The use of a redescending ψ function gives zero weight for the values of X above a certain tuning constant, so extremely large outliers do not enter the function. For this function,

$$K_p = \frac{p}{2}\chi^2_{(p+2;c_\circ^2)} - \frac{p(p+2)}{2c_\circ^2}\chi^2_{(p+4;c_\circ^2)} + \frac{p(p+2)(p+4)}{6c_\circ^4}\chi^2_{(p+6;c_\circ^2)} + \frac{c_\circ^2}{6}[1 - \chi^2_{(p;c_\circ^2)}], \tag{4}$$

where $\chi^2_{(p;c_\circ^2)}$ denotes the cumulative distribution for a χ^2 variable on p degrees of freedom, evaluated at $c_\circ^2$. Define r such that $0 < K_p/\rho(c_\circ) = r \leqslant (n-p)/2n$. For $r = (n-p)/2n$, the maximal breakdown point is $\lfloor (n-p+1)/2 \rfloor / n$ or asymptotically 0.5. As compared to the usual multivariate S-estimators which use the constraint that the average of some function ρ of the Mahalanobis distances is a constant K_p, our restriction is related to some weighted function ρ, where the weights are the cluster probabilities. The explicit expressions for the S-estimators of the cluster mean and pooled covariance matrix in the mixture model are

$$\hat{\mu_{jr}} = \frac{\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})x_{ij}}{\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})}, \tag{5}$$

$$\hat{\Sigma} = \frac{\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})(x_{ij}-\mu_{jr})(x_{ij}-\mu_{jr})'}{\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)\psi(D_{ij}^{\frac{1}{2}})D_{ij}^{\frac{1}{2}}}. \tag{6}$$

For details of derivation of Eqs. (5) and (6), see Appendix A.

2.1. Discriminant rule and S-estimators

The discriminant rule (2) based on the Bayes' posterior probabilities depends on the prior probabilities as well as the conditional distributions of $P(x\,|\,j)$. Now the S-estimators of the parameters of the multivariate location vectors and the scatter matrix would be used in the estimation of posterior probabilities $P(x\,|\,j)$.

3. Simulation

3.1. Simulation 1

For discrimination purposes, unstructured data were generated from a mixture of multivariate normal distributions. 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 $(1, \ldots, 5) \times (1, \ldots, 5)$. Each subclass was comprising of 20 observations, with a total of 240 observations in the training sample; see, for example, [3]. In this simulation, the purpose was to assess the effect of outliers on the apparent error rates, i.e., how the presence of outliers in the training data affects the classification of observations. Simulations were run with the stated configuration of unstructured data which was effectively the same for each simulation, but specifying different values of r, the theoretical breakdown point that assists in the choice of the tuning constant $c_\circ$ by satisfying $K_p/\rho(c_\circ) = r$. The errors of misclassification for both the methods are recorded in Table 1. In all the tables, the values are the average probability of misclassification averaged over 50 simulations, while the italicized values are the standard error of average probability of misclassification. For small values of r, though there were no outliers in the training data, it is apparent from Table 1 that the performance of the S-estimators was almost the same as the mda approach using the maximum likelihood estimators. It is also clear from Table 1 that as r increases, the probability of misclassification using S-estimators also increases slightly, as is to be expected. The mda approach resulted in the smallest errors of misclassification. It is because the mda approach with maximum likelihood estimators works well within the set of assumptions on which it is based.

3.2. Simulation 2

This simulation was carried out using a mixture of two bivariate Gaussian components. Each of the two groups consisted of two bivariate Gaussian subgroups.

Table 1
Errors for uncontaminated data

r	S
0.05	0.0266(*0.0094*)[a]
0.10	0.0273(*0.0104*)
0.15	0.0277(*0.0107*)
0.20	0.0281(*0.0109*)
0.25	0.0285(*0.0107*)
0.30	0.0290(*0.0107*)
mda	0.0260(*0.0098*)

[a] Average probability of misclassification averaged over 50 simulations, with standard error of average probability of misclassification in parentheses.

Table 2
Errors of misclassification for the test data

r	mda	S
0.05	0.2704(*0.0361*)[a]	0.2714(*0.0358*)
0.10	0.2546(*0.0272*)	0.2552(*0.0271*)
0.15	0.2882(*0.0402*)	0.2746(*0.0391*)
0.20	0.2635(*0.0497*)	0.2583(*0.0333*)
0.25	0.2982(*0.0809*)	0.2709(*0.0426*)
0.30	0.3087(*0.0868*)	0.2919(*0.0527*)

[a]Average probability of misclassification averaged over 50 simulations, while the italicized values in parentheses are the standard error of average probability of misclassification.

The four subgroup means were: $\mu_{11} = [1.5\ 1.5]'$, $\mu_{12} = [-1.5\ 1.5]'$, $\mu_{21} = [1.5\ -1.5]'$, $\mu_{22} = [-1.5\ -1.5]'$. 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 efficiency of the high breakdown mixture discriminant analysis was compared with the standard mda approach as in [4] using errors of misclassification. In all the tables, the values are the average probability of misclassification averaged over 50 simulations, while the italicized values are the standard error of average probability of misclassification.

Case A: The outliers generated from different bivariate Gaussians with different mean vectors, but common covariance being the positive-definite symmetric matrix were added to the training data. The cut-off point $c_\circ$ was computed for a fixed proportion of the outliers such that $K_p/\rho(c_\circ) = r$. The errors of misclassification for the test data are presented in Table 2. It is clear from Table 2 that for a higher proportion of outliers, the errors are smaller using the robust estimators as compared to the regular likelihood estimators. For 25% outliers, there is quite a significant margin in the errors of the two methods. However, it is important to mention that although an observation atypical of each component of the mixture is still clustered into one of the components, so it's contribution to the mixing proportion π_{jr} is not diminished.

Case B: In this simulation, the outliers were generated similar to case A, but with different common covariance matrix for the outliers distribution. The errors of misclassification for this simulation are presented in Table 3. It is clear that the errors using the robust estimators are smaller at higher proportion of outliers, as compared to the regular mda. For $r = 0.25$, we get 2.2% improvement in the classification using the S-estimators (see Table 3).

3.3. Simulation 3: mixtures of multivariate t-distributions

The data for the assessment of the invalidity of assumptions of the mixture of multivariate normal model were generated from mixtures of multivariate t-distributions. There were four groups and each group was a mixture of three

Table 3
Errors of misclassification, test data

r	mda	S
0.05	0.2704(*0.0341*)[a]	0.2714(*0.0359*)
0.10	0.2531(*0.0269*)	0.2557(*0.0259*)
0.15	0.2825(*0.0499*)	0.2725(*0.0414*)
0.20	0.2678(*0.0671*)	0.2557(*0.0378*)
0.25	0.2924(*0.0818*)	0.2704(*0.0383*)
0.30	0.3087(*0.0831*)	0.2903(*0.0529*)

[a] Average probability of misclassification averaged over 50 simulations, with standard error of average probability of misclassification in parentheses.

Table 4
Errors of misclassification, mixtures of multivariate-t

df	mda	S
5	0.0692(*0.0734*)[a]	0.0411(*0.0605*)
10	0.0475(*0.0648*)	0.0299(*0.0516*)
15	0.0561(*0.0818*)	0.0531(*0.0817*)
20	0.0399(*0.0677*)	0.0400(*0.0692*)
25	0.0478(*0.0717*)	0.0478(*0.0731*)
30	0.0371(*0.0695*)	0.0378(*0.0723*)

[a] Average probability of misclassification averaged over 50 simulations, with standard error of average probability of misclassification in parentheses.

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 $(1, \ldots, 5) \times (1, \ldots, 5)$. Each subclass was comprising of 20 observations, with a total of 240 observations in the training sample. The degrees of freedom for the components of the mixture model that provide a framework for assessing the degree of robustness to be incorporated into the fitting of the mixture model were varied. The test data were of size 120, with equal priors for the four groups. The errors of misclassification using the standard mda approach and the robust S-estimation approach are recorded in Table 4. However, as shown in this table the errors of misclassifications are smaller for the robust procedures as compared to the mda at smaller degrees of freedom. As the degrees of freedom for the components of the mixture increase, the errors of misclassification using both the methods decrease. With degrees of freedom approaching to 20, the errors of misclassification using the robust approach are almost the same as from the standard mda approach. It is due to the fact that the t-distribution approaches the normal for larger degrees of freedom. So, the standard mda approach based on the maximum likelihood method performed better, because the distributional assumption was satisfied in this case.

4. Conclusion

This paper introduces the idea of high breakdown discrimination, when the distributions are mixtures of multivariate normal. The robust S-estimators performed well in our simulation study, when the data were either contaminated or the assumption of multivariate normality was invalid. However, the classification performance based on S-estimators is only marginally improved as compared to mda because the effect of outliers on the mixing proportions is not diminished. The S-estimators being much faster redescending have been shown to behave better for the extreme outlier configuration in the simulation studies. But in the case of mixtures of subgroups, it is difficult to assess whether an outlier is a true outlier or an outlier for one subgroup might be an inlier for the other. Further, by initial clustering, the outliers are still clustered into one of the subgroups, affecting the initial estimates, as well as the cluster probability which is not robustified. This is another cause of the slight improvement in classification by using the S-estimators. However, further research can be conducted to explore this area.

We have shown that the use of high breakdown point estimators help in the improvement of performance of mixture discriminant analysis in the presence of contaminated data. The robust S-estimators also help in the choice of smaller number of subgroups, by identifying a group of outliers which are in case of mda approach clustered as a separate subgroup.

Appendix A

Proof of Eqs. (5) and (6).

Using Lagrange multipliers for the minimization problem in case of S-estimators as defined in Definition 1

$$L = \log|\tilde{\Sigma}| - \lambda \times \left[\frac{1}{n}\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)\rho\{(x_{ij}-\mu_{jr})'\Sigma^{-1}(x_{ij}-\mu_{jr})\}^{1/2} - K_p\right]. \tag{A.1}$$

The first-order conditions from the minimization problem of Eq. (A.1) are $\partial L/\partial\mu_{jr} = 0$, or

$$\frac{\lambda}{n}\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})\Sigma^{-1}(x_{ij}-\mu_{jr}) = 0, \tag{A.2}$$

where $u(D_{ij}) = \psi(D_{ij})/D_{ij}$. Taking derivative of Eq. (A.1) with respect to Σ, i.e., $\partial L/\partial\Sigma = 0$, we obtain

$$\Sigma^{-1} - \frac{\lambda}{2n}\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j).\frac{\rho'\{(x_{ij}-\mu_{jr})'\Sigma^{-1}(x_{ij}-\mu_{jr})\}^{1/2}}{\{(x_{ij}-\mu_{jr})'\Sigma^{-1}(x_{ij}-\mu_{jr})\}^{1/2}} \times \{-2\Sigma^{-1}(x_{ij}-\mu_{jr})(x_{ij}-\mu_{jr})'\Sigma^{-1} + d_{\Sigma^{-1}(x_{ij}-\mu_{jr})(x_{ij}-\mu_{jr})'\Sigma^{-1}}\} = 0,$$

where d is a $p \times p$ diagonal matrix, with *ith* element of d as the *ith* diagonal element of aa'. On simplification,

$$I + \frac{\lambda}{2n}\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})A^{-1}(x_{ij}-\mu_{jr})(x_{ij}-\mu_{jr})'A^{-T} = 0. \quad \text{(A.3)}$$

Taking the trace yields

$$p + \frac{\lambda}{2n}\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)\psi(D_{ij}^{\frac{1}{2}})D_{ij}^{\frac{1}{2}} = 0.$$

Solving we obtain,

$$\lambda = -\frac{2np}{\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)\psi(D_{ij}^{\frac{1}{2}})D_{ij}^{\frac{1}{2}}}. \quad \text{(A.4)}$$

Now substituting this value of λ in (A.2) and (A.3), gives the estimates of the cluster mean and covariance matrix as

$$\hat{\mu_{jr}} = \frac{\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})x_{ij}}{\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})}.$$

$$\hat{\Sigma} = \frac{\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)u(D_{ij}^{\frac{1}{2}})(x_{ij}-\mu_{jr})(x_{ij}-\mu_{jr})'}{\sum_{j=1}^{J}\sum_{r=1}^{R_j}\sum_{i=1}^{n_j} P(C_{jr}|x_{ij},j)\psi(D_{ij}^{\frac{1}{2}})D_{ij}^{\frac{1}{2}}}.$$

References

[1] C. Croux, C. Dehon, Robust linear discriminant analysis using *S*-estimators, Canad. J. Statist. 29 (2001) 473–493.

[2] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data, J. Roy. Statist. Soc. Ser. B 39 (1997) 1–38.

[3] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, Technical Report, AT&T Bell Laboratories, Murray Hill, 1994.

[4] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. Roy. Statist. Soc. Ser. B 58 (1) (1996) 155–176.
[5] D. Hawkins, Identification of Outliers, Chapman & Hall, London, 1980.
[6] X. He, W.K. Fung, High breakdown estimation for multiple populations with applications to discriminant analysis, J. Multivariate Anal. 72 (2000) 151–162.
[7] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[8] H.P. Lopuhaa, On the relation between S-estimators and M-estimators of multivariate location and covariance, Ann. Statist. 17 (1989) 1662–1683.
[9] R.A. Maronna, Robust M estimators of multivariate location and scatter, Ann. Statist. 4 (1976) 51–67.
[10] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.
[11] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Applications to Clustering, Marcel Dekker, New York, 1988.
[12] P.J. Rousseeuw, Least median of squares regression, J. Amer. Statist. Assoc. 79 (1984) 871–880.
[13] P.J. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[14] D.M. Titterington, Some recent research in the analysis of mixture distributions, Statistics 21 (1990) 619–641.