UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Classification and Discrimination for populations with Mixture of Multivariate Normal Distributions

Emerson Wruck
Jorge Alberto Achcar
Josmar Mazucheli

Nº 62

NOTAS

São Carlos - SP

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2577

Classification and Discrimination for populations with Mixture of Multivariate Normal Distributions

Emerson Wruck
Jorge Alberto Achcar
Josmar Mazucheli

Nº 62

NOTAS

Série Estatística

São Carlos – SP
Out./2000

CLASSIFICATION AND DISCRIMINATION FOR POPULATIONS WITH MIXTURE OF MULTIVARIATE NORMAL DISTRIBUTIONS

Emerson Wruck *

Jorge Alberto Achcar

ICMC, Universidade de São Paulo, C.P. 668, 13560-970, São Carlos, S.P., Brazil

Josmar Mazucheli

DEs, Universidade Estadual de Maringá, Maringá,P.R.,Brazil

■ RESUMO: Neste artigo, consideramos o uso de misturas de distribuições normais multivariadas para serem usadas em leis de classificação e discriminação. Considerando o uso de métodos de Monte Carlo em Cadeias de Markov, obtemos sumários a posteriori de interesse e densidades preditivas para serem usadas nas leis de classificação. Um exemplo é introduzido.

■ ABSTRACT: In this paper, we consider the use of mixtures of multivariate normal distributions to be used in classification and discrimination rules. Considering the use of Markov Chain Monte Carlo methods, we get the posterior summaries of interest and the predictive densities to be used in the classification rules. A numerical example is introduced to illustrate the proposed methodology.

■ KEYWORDS: Mixture of multivariate normal distributions, classification and discrimination, Bayesian analysis.

1 Introduction

Let us assume that we have interest to classify a unit to one among g groups based on a vetor x of observed data (see for example, Cacoulos,1973; Lachenbruch,1975; Goldstein

[1*] O primeiro autor agradece o apoio financeiro da FAPESP, processo ♯ 99/08961-0

and Dillon, 1978; or Johnson and Wichern, 1982). This problem usually appears in different areas, as economy, medicine, ecology, archeology or physics.

Usually, x is assumed to have a multivariate normal distribution (see for example, Anderson,1984; or Johnson and Wichern, 1982). The classification rule could be based on Fisher's discriminant function or using Bayesian approaches based on the predictive density for a future observation (see for example, Lavine and West, 1992).

For many applications, a preliminary data analysis for a trainning data set of the g populations could indicate the need for other multivariate distributions for x, which could improve the performance of the classification rules. Evaluation of classification rules could be based on "error rates"or misclassifications probabilities.

In this paper, we assume a mixture of multivariate distributions for x in each population with density,

$$f(\mathrm{x}|\boldsymbol{\theta}) = \sum_{j=1}^{K} p_j f(\mathrm{x}|\boldsymbol{\theta}_j) \tag{1}$$

where $\boldsymbol{\theta}_j$ is a vector of parameters associated to the jth component distribution and $\sum_{j=1}^{K} p_j = 1$.

Bayesian inference for mixture distributions is introduced by many authors (see for exemple, Robert, 1996; or Titterington, Smith and Markov,1985).

As a special case, we consider a Bayesian approach for classification assuming a mixture of multivariate normal distributions for each population using MCMC (Markov Chain Monte Carlo) methods as in Gelfand and Smith (1990) to develop the classification rules.

2 Bayesian Analysis Assuming a Mixture of K=2 Multivariate Normal Distributions

First of all, we assume the special case where each population have a mixture of K=2 multivariate normal distributions,

$$f(\mathrm{x}|\boldsymbol{\theta}) = \sum_{j=1}^{2} p_j f(\mathrm{x}|\boldsymbol{\theta}_j) \tag{2}$$

where $\mathrm{x} = (x_1, ..., x_q)$; $\sum_{j=1}^{2} p_j = 1$; $f_j(\mathrm{x}|\boldsymbol{\theta}_j)$ denotes a multivariate normal distribution $N_q(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$; $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$; $\boldsymbol{\theta}_1 = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}$ and $\boldsymbol{\theta}_2 = \{\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\}$.

The likelihood function for $\boldsymbol{\theta}$ and $\mathrm{p} = (p_1, p_2)$ based on a random sample $\mathrm{X}_1, ..., \mathrm{X}_n$ is given by,

$$L(\boldsymbol{\theta}, \mathrm{p}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{2} p_j f_j(\mathrm{x}_i|\boldsymbol{\theta}_j) \right\} \tag{3}$$

For simplification of the conditional distributions needed for the Gibbs Sampling algorithm, we introduce latent variables (see Tanner and Wong, 1987), $\mathrm{Z}_i = (Z_{i1}, Z_{i2})$ where $Z_{ij} = 1$ if the ith observation was generated from the jth component distribution ($Z_{ij} = 0$ in other case) and $\sum_{j=1}^{2} Z_{ij} = 1, \ i = 1, ..., n$.

Observe that for the special case of $K = 2$ component distributions, $Z_{ij}|\mathrm{x}, \boldsymbol{\theta}, \mathrm{p} \sim b(1, v_{ij})$ (a binomial distribution) with

$$v_{ij} = \frac{p_j f_j(\mathrm{x}_i|\boldsymbol{\theta}_j)}{\sum_{j=1}^{2} p_j f_j(\mathrm{x}_i|\boldsymbol{\theta}_j)} \tag{4}$$

Thus,

$$f(\mathrm{z}_i|\mathrm{x}, \boldsymbol{\theta}, \mathrm{p}) = v_{i1}^{z_{i1}} (1 - v_{i1})^{1 - z_{i1}} \tag{5}$$

Considering a sample $\mathrm{Z}_1, ..., \mathrm{Z}_n$, we have,

$$f(\mathrm{z}_1, ..., \mathrm{z}_n|\mathrm{x}, \boldsymbol{\theta}, \mathrm{p}) = \frac{\prod_{i=1}^{n} \prod_{j=1}^{2} [p_j f(\mathrm{x}_i|\boldsymbol{\theta}_j)]^{z_{ij}}}{\prod_{i=1}^{n} \left\{ \sum_{j=1}^{2} p_j f(\mathrm{x}_i|\boldsymbol{\theta}_j) \right\}} \tag{6}$$

Let us assume the following prior distribution for $\boldsymbol{\theta}$ and p_1 (see for example, Lavine and West, 1992):

$$\begin{aligned} &\pi(\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}_1|^{-\frac{1}{2}(q+1)} |\boldsymbol{\Sigma}_2|^{-\frac{1}{2}(q+1)} \\ &\pi(p_1) \sim B(a, b); \ a, b \, \text{known}. \end{aligned} \tag{7}$$

where $B(a, b)$ denotes a Beta distribution with mean $\frac{a}{(a+b)}$ and variance $\frac{ab}{[(a+b)^2(a+b+1)]}$.

Combining (3) with (6) and the prior distribution (7) assuming independence, the joint posterior distribution for $\boldsymbol{\theta}$ and p_1 is given by

$$\begin{aligned} \pi(\boldsymbol{\theta}, p_1|\mathrm{x}, \mathrm{z}) \propto |\boldsymbol{\Sigma}_1|^{-\frac{1}{2}(q+1)} \prod_{i=1}^{n} \left\{ [f_1(\mathrm{x}_i|\boldsymbol{\theta}_1)]^{z_{i1}} \right\} |\boldsymbol{\Sigma}_2|^{-\frac{1}{2}(q+1)} \prod_{i=1}^{n} \left\{ [f_2(\mathrm{x}_i|\boldsymbol{\theta}_2)]^{z_{i2}} \right\} \\ p_1^{(r+a)-1} (1 - p_1)^{(n+b-r)-1} \end{aligned} \tag{8}$$

where $r = \sum_{i=1}^{n} z_{i1}$ e $r_2 = n - r = \sum_{i=1}^{n} z_{i2}$.

The conditional posterior distributions for the Gibbs Sampling algorithm are given by,

$$
\begin{aligned}
&(i) \quad p_1|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathrm{x}, \mathrm{z} \sim B(a+r; b+n-r); \\
&(ii) \quad \boldsymbol{\mu}_1|\boldsymbol{\Sigma}_1, p_1, \boldsymbol{\theta}_2, \mathrm{x}, \mathrm{z} \sim N_q(\bar{\mathrm{x}}_1; \tfrac{1}{r}\boldsymbol{\Sigma}_1); \\
&(iii) \quad \boldsymbol{\Sigma}_1|\boldsymbol{\mu}_1, p_1, \boldsymbol{\theta}_2, \mathrm{x}, \mathrm{z} \sim Inv - Wishart_{r-1}(V_1^{-1}); \\
&(iv) \quad \boldsymbol{\mu}_2|\boldsymbol{\Sigma}_2, p_1, \boldsymbol{\theta}_1, \mathrm{x}, \mathrm{z} \sim N_q(\bar{\mathrm{x}}_2; \tfrac{1}{n-r}\boldsymbol{\Sigma}_2); \\
&(v) \quad \boldsymbol{\Sigma}_2|\boldsymbol{\mu}_2, p_1, \boldsymbol{\theta}_1, \mathrm{x}, \mathrm{z} \sim Inv - Wishart_{n-r-1}(V_2^{-1});
\end{aligned}
\tag{9}
$$

where $\bar{x}_1 = \frac{1}{r}\sum_{i=1}^{n} Z_{i1}\mathrm{x}_i$; $\bar{x}_2 = \frac{1}{r_2}\sum_{i=1}^{n} Z_{i2}\mathrm{x}_i$; $\mathbf{V}_1 = \sum_{i=1}^{n} Z_{i1}(\mathrm{x}_i - \bar{x}_1)(\mathrm{x}_i - \bar{x}_1)'$; $\mathbf{V}_2 = \sum_{i=1}^{n} Z_{i2}(\mathrm{x}_i - \bar{x}_2)(\mathrm{x}_i - \bar{x}_2)'$ and $Inv - Wishart_v(S^{-1})$ denotes a Inverse-Wishart distribution with v degrees of freedom with density

$$
f(W) \propto |W|^{-\frac{1}{2}(v+q+1)} \exp\left\{-\frac{1}{2}tr|SW^{-1}|\right\},
$$

S is a $q \times q$ symetric positive-definide scale matrix and W is positive-definide.

To generate samples from the joint posterior distribution (8), we follow the steps:

i- Start with initial values $p_1^{(0)}, \boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \boldsymbol{\Sigma}_1^{(0)}$ and $\boldsymbol{\Sigma}_2^{(0)}$;

ii- Generate a sample $\mathrm{Z}_1^{(1)}, ..., \mathrm{Z}_n^{(1)}$ from a binomial distribution with success probability v_{ij} (4).

iii- Generate a sample of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ from the conditional distributions (9).

We also could consider a informative prior distribution for $\boldsymbol{\theta}$. A conjugate prior distribution for $\boldsymbol{\theta}$ is given by,

$$
\begin{aligned}
\pi(\boldsymbol{\theta}) = &|\boldsymbol{\Sigma}_1|^{-(\frac{(g_1+q)}{2}+1)} \exp\left\{-\frac{1}{2}tr[\mathrm{G}_1\boldsymbol{\Sigma}_1^{-1}] - \frac{k_1}{2}(\boldsymbol{\mu}_1 - \mathrm{m}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \mathrm{m}_1)\right\} \\
&|\boldsymbol{\Sigma}_2|^{-(\frac{(g_2+q)}{2}+1)} \exp\left\{-\frac{1}{2}tr[\mathrm{G}_2\boldsymbol{\Sigma}_2^{-1}] - \frac{k_2}{2}(\boldsymbol{\mu}_2 - \mathrm{m}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \mathrm{m}_2)\right\}
\end{aligned}
\tag{10}
$$

where g_j and k_j are known constants; G_j is a matrix of known constants; m_j is a vector of known constants, $j = 1, 2$.

With prior (10) for $\boldsymbol{\theta}$ and the same Beta prior for p_1 given in (7), the conditional posterior distributions for the Gibbs algorithm are given by,

$$
\begin{aligned}
&(i) \quad p_1|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathrm{x}, \mathrm{z} \sim B(a+r; b+n-r); \\
&(ii) \quad \boldsymbol{\mu}_1|\boldsymbol{\Sigma}_1, p_1, \boldsymbol{\theta}_2, \mathrm{x}, \mathrm{z} \sim N_q(\mathrm{a}_1; \tfrac{\boldsymbol{\Sigma}_1}{r+k_1}); \\
&(iii) \quad \boldsymbol{\Sigma}_1|\boldsymbol{\mu}_1, p_1, \boldsymbol{\theta}_2, \mathrm{x}, \mathrm{z} \sim Inv-Wishart_{g_1+r}(\mathbf{G}_{n_1}^{-1}); \\
&(iv) \quad \boldsymbol{\mu}_2|\boldsymbol{\Sigma}_2, p_1, \boldsymbol{\theta}_1, \mathrm{x}, \mathrm{z} \sim N_q(\mathrm{a}_2; \tfrac{\boldsymbol{\Sigma}_2}{n-r+k_2}); \\
&(v) \quad \boldsymbol{\Sigma}_2|\boldsymbol{\mu}_2, p_1, \boldsymbol{\theta}_1, \mathrm{x}, \mathrm{z} \sim Inv-Wishart_{g_2+n-r}(\mathbf{G}_{n_2}^{-1});
\end{aligned}
\tag{11}
$$

where $\mathrm{a}_1 = \frac{r}{r+k_1}\bar{\mathrm{x}}_1 + \frac{k_1}{r+k_1}\mathrm{m}_1$; $\mathrm{a}_2 = \frac{n-r}{n-r+k_2}\bar{\mathrm{x}}_2 + \frac{k_2}{n-r+k_2}\mathrm{m}_2$; $\mathbf{G}_{n1}^{-1} = \mathbf{G}_1 + \mathbf{V}_1 + \left(\frac{k_1 r}{k_1+r}\right)(\bar{\mathrm{x}}_1 - \mathrm{m}_1)(\bar{\mathrm{x}}_1 - \mathrm{m}_1)'$ and $\mathbf{G}_{n2}^{-1} = \mathbf{G}_2 + \mathbf{V}_2 + \left(\frac{k_2(n-r)}{k_2+n-r}\right)(\bar{\mathrm{x}}_2 - \mathrm{m}_2)(\bar{\mathrm{x}}_2 - \mathrm{m}_2)'$.

Similar results could be obtained considering $K > 2$.

3 Classification for two Populations

Let us classify a new object to one of two populations based on q measurements associated random variables $\mathrm{X}' = (X_1, ..., X_q)$ assuming a mixture of normal distributions $f^{(1)}(\mathrm{x}|\boldsymbol{\theta}^{(1)}) = \sum_{j=1}^{2} p_j^{(1)} f_j^{(1)}(\mathrm{x}|\boldsymbol{\theta}_j^{(1)})$ for population 1 and $f^{(2)}(\mathrm{x}|\boldsymbol{\theta}^{(2)}) = \sum_{j=1}^{2} p_j^{(2)} f_j^{(2)}(\mathrm{x}|\boldsymbol{\theta}_j^{(2)})$ for population 2 where $\boldsymbol{\theta}_j^{(l)} = (\mu_j^{(l)}, \boldsymbol{\Sigma}_j^{(l)})$ and $f_j^{(l)}(\mathrm{x}|\boldsymbol{\theta}_j^{(l)})$ denotes a multivariate normal distribution $N_q(\mu_j^{(l)}; \boldsymbol{\Sigma}_j^{(l)})$; $j = 1, 2$; $l = 1, 2$.

The predictive density for a vector x is given by,

$$
f^{(l)}(\mathrm{x}) = \int f^{(l)}(\mathrm{x}|\boldsymbol{\theta}^{(l)}) \pi(\boldsymbol{\theta}^{(l)}|\mathrm{x}) d\boldsymbol{\theta}^{(l)} \tag{12}
$$

where $l = 1$ or 2 (l indexes populations 1 and 2).

A Monte Carlo estimate for $f^{(l)}(\mathrm{x})$ based on the generated Gibbs Samples is given by,

$$
\hat{f}^{(l)}(\mathrm{x}) = \frac{1}{S} \sum_{s=1}^{S} f^{(l)}(\mathrm{x}|\boldsymbol{\theta}^{(l)s}) \tag{13}
$$

where S is the number of generated Gibbs Samples.

To classify a new object with observed measurements x, we consider the following allocation rule:

i- Allocate x to population 1 if,

$$\frac{\hat{f}^{(1)}(\mathrm{x})}{\hat{f}^{(2)}(\mathrm{x})} \geq \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right] \tag{14}$$

where $c(1|2)$ is the missclassification cost when an observation from population 2 is incorrectly classified in population 1; $c(2|1)$ is the missclassification cost when an observation from population 1 is incorrectly classified in population 2; p_1 and p_2 are the prior probabilities of classification to populations 1 and 2, respectively.

ii- Allocate x to population 2, otherwise.

In the special case of $c(1|2) = c(2|1)$ and $p_1 = p_2$, the allocation rule (14) is given by,

i- Allocate x to population 1 if,

$$\frac{\hat{f}^{(1)}(\mathrm{x})}{\hat{f}^{(2)}(\mathrm{x})} \geq 1 \tag{15}$$

ii- Allocate x to population 2, otherwise.

4 A Numerical Illustration

As an illustrative example, let us consider the data of two simulate samples of size 100 generated from populations 1 and 2 with mixtures of two bivariate normal distribution with density (2).

For population 1, we assume,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2.5 & 4.5 \end{pmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1.5 \end{pmatrix}; \quad p_1 = 0.4$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} 4.0 & 10.0 \end{pmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.0 & 0.4 \\ 0.4 & 2.5 \end{pmatrix}; \quad p_2 = 0.6$$

For population 2, we assume,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 3.5 & 5.5 \end{pmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.0 & 0.3 \\ 0.3 & 2.0 \end{pmatrix}; \quad p_1 = 0.5$$

$$\mu_2 = \begin{pmatrix} 6.5 & 14.6 \end{pmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2.0 & 0.4 \\ 0.4 & 3.0 \end{pmatrix}; \quad p_2 = 0.5$$

In figure 1, we have the plot for the data $\mathbf{x} = (x_1, x_2)$ from both populations.

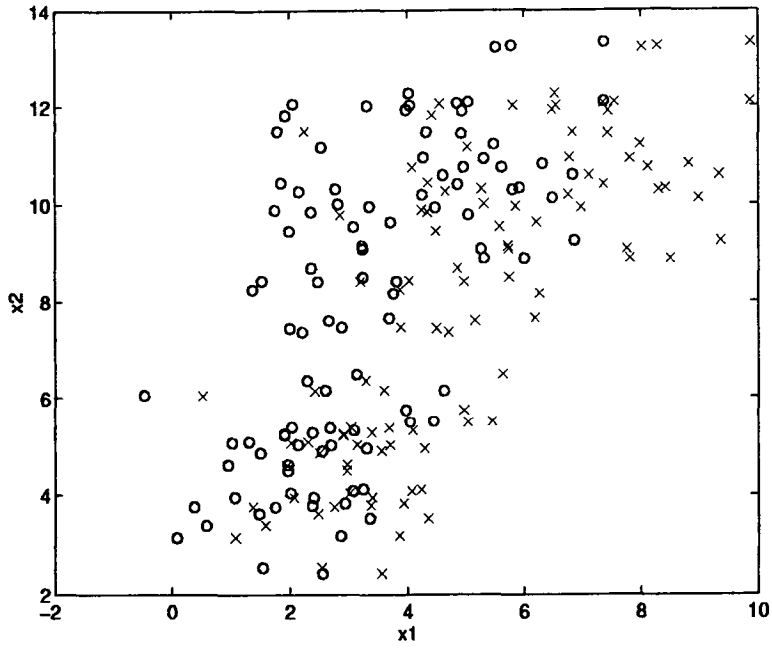

Figure 1: Data from populations 1(o) and 2 (×).

In figure 1, we clearly observe two clusters of data for both samples, which indicates the mixture of two bivariate normal distributions for each sample.

If we consider the usual linear discriminant function assuming multivariate normal distributions for each population with same covariance matrix Σ (see for example, Johnson and Wichern, 1982), same missclassification costs and same prior probabilities, we have in table 1, the classification table for all data set.

Table 1 - Classification table (linear discriminant function)

	Predicted membership			
		Pop1	Pop2	Total
Actual	Pop1	n_{1c} =70	n_{1m} = 30	n_1 = 100
membership	Pop2	n_{2m} =40	n_{2c} = 60	n_2 =100

In table 1, n_{1c} is the number of Pop1 items correctly classified as Pop1 items; n_{1m} is the number of Pop1 items missclassified as Pop2 items; n_{2c} is the number of Pop2 items

correctly classified; n_{2m} is the number of Pop2 items missclassified; n_1 and n_2 are the total of actual items in each population.

The apparent error rate is given by

$$APER = \frac{n_{1m} + n_{2m}}{n_1 + n_2} = \frac{30 + 40}{100 + 100} = 0.35.$$

Considering a quadratic discriminant function assuming multivariate normal distributions for each population with different covariance matrices $\Sigma_1 \neq \Sigma_2$, same missclassification costs and same prior probabilities, we have in table 2, the classification table for all data set.

Table 2 - Classification table (quadratic discriminant function)

	Predicted membership			
		Pop1	Pop2	Total
Actual	Pop1	n_{1c} =76	n_{1m} = 24	n_1 = 100
membership	Pop2	n_{2m} =41	n_{2c} = 59	n_2 =100

Using the quadratic discriminant function, the apparent error rate is given by $APER = (24 + 41)/200 = 0.325$.

We observe from the values for $APER$ considering a usual linear discriminant function or a quadratic discriminant function that a large proportion of items are missclassified, which indicates that the used classification rules are not appropriated for the data set.

Considering a mixture of two bivariate normal distributions (2) with $\boldsymbol{\theta}_1^{(l)} = (\boldsymbol{\mu}_1^{(l)}, \boldsymbol{\Sigma}_1^{(l)})$ and $\boldsymbol{\theta}_2^{(l)} = (\boldsymbol{\mu}_2^{(l)}, \boldsymbol{\Sigma}_2^{(l)})$ where,

$$\boldsymbol{\mu}_1^{(l)} = (\mu_{11}^{(l)}, \mu_{12}^{(l)}), \quad \boldsymbol{\mu}_2^{(l)} = (\mu_{21}^{(l)}, \mu_{22}^{(l)}), \quad \boldsymbol{\Sigma}_1^{(l)} = \begin{pmatrix} \sigma_{111}^{(l)} & \sigma_{112}^{(l)} \\ \sigma_{121}^{(l)} & \sigma_{122}^{(l)} \end{pmatrix}, \quad \boldsymbol{\Sigma}_2^{(l)} = \begin{pmatrix} \sigma_{211}^{(l)} & \sigma_{212}^{(l)} \\ \sigma_{221}^{(l)} & \sigma_{222}^{(l)} \end{pmatrix}$$

for $l = 1$ (Pop1) and $l = 2$ (Pop2) and the prior distributions (7) with $a = 2, b = 3$ for Pop1 and $a = 2, b = 2$ for Pop2, we generated 10000 Gibbs Samples for the joint posterior distribution (8) using the conditional posterior distributions (9). We monitored the convergence of the Gibbs Samples using the Geweke (1992) method.

The results were generated using Ox package version 2.10 (see Doornik,1999). For each parameter, we discarded the 4000 first iterations (**burn-in-samples**) and we considered the 20th,40th,... iterations. Therefore, we have a final sample of size $S = 300$.

In table 3, we have the posterior summaries for all parameters. We also have in table 3, the convergence values for the Geweke (1992) criterium GW. We observe convergence for all parameters since $|GW| < 2$.

Considering Monte Carlo estimates for the predictive densities for x in both populations based on the $S = 300$ generated Gibbs Samples, we use (15) to classify the items to both populations 1 and 2.

Table 3 - Posterior summaries (mixture of two bivariate normal distributions; prior distributions (7) for $\boldsymbol{\theta}$)

	Parameter	Mean	S.D.	95% credible interval	$\mid GW \mid$
Pop1	$\mu_{11}^{(1)}$	2.2180	0.20796	(1.8231;2.5952)	0.0187
	$\mu_{12}^{(1)}$	4.7413	0.36945	(4.1906;5.6530)	0.0312
	$\sigma_{111}^{(1)}$	1.2288	0.33482	(0.70499;2.0119)	0.0256
	$\sigma_{122}^{(1)}$	2.0556	1.1650	(0.79698;5.2064)	0.0104
	$\sigma_{112}^{(1)}$	0.38458	0.32913	(-0.22956;1.0659)	0.0864
	$p_1^{(1)}$	0.38848	0.069225	(0.26940;0.54735)	0.0455
	$\mu_{21}^{(1)}$	4.0117	0.28724	(3.4918;4.7025)	0.0163
	$\mu_{22}^{(1)}$	10.0520	0.40065	(9.1946;10.722)	0.0523
	$\sigma_{211}^{(1)}$	2.5543	0.50716	(1.6696;3.7523)	0.0164
	$\sigma_{222}^{(1)}$	4.1283	1.4710	(2.0013;8.0003)	0.0704
	$\sigma_{212}^{(1)}$	1.1125	0.5469	(0.15604;2.3419)	0.0175
Pop2	$\mu_{11}^{(2)}$	3.1585	0.1700	(2.8474;3.5063)	0.0497
	$\mu_{12}^{(2)}$	5.5554	0.18671	(5.1742;5.9197)	0.0539
	$\sigma_{111}^{(2)}$	1.2615	0.28094	(0.82515;1.9301)	0.0025
	$\sigma_{122}^{(2)}$	1.6764	0.43747	(1.0427;2.5983)	0.0131
	$\sigma_{112}^{(2)}$	0.18441	0.2566	(-0.31451;0.7831)	0.0327
	$p_1^{(2)}$	0.44199	0.046343	(0.3554;0.52604)	0.0075
	$\mu_{21}^{(2)}$	6.5406	0.2432	(5.9708;7.0236)	0.0036
	$\mu_{22}^{(2)}$	14.80	0.2556	(14.250;15.305)	0.0310
	$\sigma_{211}^{(2)}$	2.4759	0.49652	(1.7055;3.5964)	0.0055
	$\sigma_{222}^{(2)}$	4.4486	1.0596	(2.7541;7.0163)	0.0112
	$\sigma_{212}^{(2)}$	1.2344	0.55629	(0.19226;2.5609)	0.0174

In table 4, we have the classification table for all data.

Table 4 - Classification table (mixture of two bivariate normal distributions.)

	Predicted membership			
		Pop1	Pop2	Total
Actual membership	Pop1	n_{1c} =83	n_{1m} = 17	n_1 = 100
	Pop2	n_{2m} =19	n_{2c} = 81	n_2 =100

Considering a mixture of two bivariate normal distributions for both populations, the apparent error rate is given by $APER = (17 + 19)/200 = 0.18$. That is, we observe a great improvement in the classification rule using the mixture of two bivariate normal distributions, since we get a very small value the $APER$ in comparison with the obtained values for the $APER$ using linear or quadratic discriminant functions.

We also could assume the conjugate prior distribution (10) for $\boldsymbol{\theta}$. Considering

$$\mathrm{m}_1 = (2.5; 4.5);\ \mathrm{m}_2 = (4.0; 10.0);\ k_1 = k_2 = 3; g_1 = g_2 = 7; a = b = 10;$$

$$\mathbf{G}_1 = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1.5 \end{pmatrix} \text{ and } \mathbf{G}_2 = \begin{pmatrix} 2.0 & 0.4 \\ 0.4 & 2.5 \end{pmatrix}$$

for population 1 and

$$\mathrm{m}_1 = (3.5; 5.5);\ \mathrm{m}_2 = (6.5; 14.6);\ k_1 = k_2 = 3; g_1 = g_2 = 7; a = b = 10;$$

$$\mathbf{G}_1 = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 2 \end{pmatrix} \text{ and } \mathbf{G}_2 = \begin{pmatrix} 2.0 & 0.4 \\ 0.4 & 3.0 \end{pmatrix}$$

for population 2, we have in table 5, the posterior summaries for all parameters based on $S = 10000$ Gibbs Samples generated from the conditional posterior distributions (11). The used simulation procedure was similar for the case considering the prior distribution (7).

Table 5 - Posterior summaries (mixture of two bivariate normal distributions; prior distributions (10) for $\boldsymbol{\theta}$)

	Parameter	Mean	S.D.	95% credible interval	$\mid GW \mid$
	$\mu_{11}^{(1)}$	2.2626	0.19234	(1.8744;2.6433)	0.0019
	$\mu_{12}^{(1)}$	4.5904	0.22286	(4.2052;5.0603)	0.0116
	$\sigma_{111}^{(1)}$	0.94689	0.2639	(0.55589;1.6046)	0.0736
	$\sigma_{122}^{(1)}$	1.3017	0.47161	(0.70885;2.5620)	0.0588
	$\sigma_{112}^{(1)}$	0.3243	0.20275	(0.004668;0.81366)	0.0719
Pop1	$p_1^{(1)}$	0.39912	0.052701	(0.29853;0.50799)	0.0189
	$\mu_{21}^{(1)}$	3.9721	0.22622	(3.5397;4.4167)	0.0531
	$\mu_{22}^{(1)}$	10.054	0.28728	(9.4443;10.579)	0.1254
	$\sigma_{211}^{(1)}$	2.233	0.39157	(1.616;3.0994)	0.0404
	$\sigma_{222}^{(1)}$	3.6273	0.95053	(2.0555;5.8916)	0.0801
	$\sigma_{212}^{(1)}$	0.99682	0.42114	(0.26592;1.9683)	0.0718
	$\mu_{11}^{(2)}$	3.2034	0.16598	(2.8589;3.5076)	0.0114
	$\mu_{12}^{(2)}$	5.6200	0.1778	(5.2761;5.9473)	0.0023
	$\sigma_{111}^{(2)}$	1.0981	0.2387	(0.71551;1.5960)	0.0348
	$\sigma_{122}^{(2)}$	1.5455	0.35837	(0.98285;2.3600)	0.0296
	$\sigma_{112}^{(2)}$	0.20325	0.19768	(-0.20371;0.63515)	0.0405
Pop2	$p_1^{(2)}$	0.45774	0.044559	(0.37268;0.53877)	0.0274
	$\mu_{21}^{(2)}$	6.5791	0.20661	(6.1442;6.9367)	0.0154
	$\mu_{22}^{(2)}$	14.811	0.23635	(14.385;15.286)	0.0009
	$\sigma_{211}^{(2)}$	2.1137	0.41496	(1.4666;3.1241)	0.0266
	$\sigma_{222}^{(2)}$	3.7948	0.72236	(2.559;5.3511)	0.0649
	$\sigma_{212}^{(2)}$	1.0023	0.41995	(0.26454;1.8881)	0.0542

In table 6, we have the classification table for all data.

Table 6 - Classification table (mixture of two bivariate normal distribution and the priori distributions (10) for $\boldsymbol{\theta}$.)

	Predicted membership		
		Pop1	Pop2	Total
Actual membership	Pop1	n_{1c} =86	n_{1m} = 14	n_1 = 100
	Pop2	n_{2m} =18	n_{2c} = 82	n_2 =100

In this case, the apparent error rate is given by $APER = (14 + 18)/200 = 0.16$.

We observe a better performance for the classification rule assuming a mixture of bivariate normal distributions for both population and the conjugate prior distribution (10).

5 Concluding Remarks

For many problems of classification and discrimination, the use of standard linear or quadratic discriminant functions could not be appropriate. Usually, a preliminary analysis of existing trainning data could indicate different shapes for the multivariate distribution to be used in the classification rules in place of the usual assumption of multivariate normal distribution for the data of each population.

In this case, the use of mixture of multivariate normal distributions could be very useful to be used in the classification rules.

It is important to point out that the use of MCMC methods to get the posterior summaries of interest does not require sophisticate computational expertise and this approach could be extended to mixtures of more than two multivariate normal distributions with higher dimensions.

References

1 ANDERSON, T. W. *An Introduction to Multivariate Statistical Methods.* New York: John Wiley, 1984.

2 CACOULLOS, T. *Discriminant Analysis and Applications.* New York: Academic Press, 1973.

3 DOORNIK,J.A.. *Object-Oriented Matrix Programming Using Ox.*,3rd ed. London: Timberlake Consultants Press and Oxford, 1999.

4 GEWEKE, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.*In: Bayesian Statistics 4, J. M. Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger (eds.),* pp. 169-193. New York: Oxford University Press.

5 GELFAND, A.; SMITH, A. Sampling based approaches to calculating marginal densities.*Journal of the American Statistical Association.*, v.85, p.398-409, 1990.

6 GOLDSTEIN, M.;DILLON, W. R. *Discriminant Analysis.* New York: Wiley,1978.

7 JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey: Prentice Hall,1982.

8 LACHENBRUCH, P. A. *Discriminant Analysis.*New York: Hafner, 1975.

9 LAVINE, M.; WEST, M. A Bayesian method for classification and discrimination.*The Canadian Journal of Statistics,*4 , v.20, pp.451-461, 1992.

10 PRESS, S. J. *Applied Multivariate Analysis.*New York: Holt,Rinehart, and Winston, 1972.

11 ROBERT, C. P.*Mixture of Distributions: Inference and Estimation, in Markov Chain Monte Carlo in Practice.* London: (Eds,W.R.Gilks, S.Richardson, D.J.Spiegelhalter), Chapman and Hall, p.441-464,1996.

12 SEBER, G. A. F. *Multivariate Observations.*New York: Wiley ,1983.

13 SMITH, A. F. M.; ROBERT, G. O. Bayesian Computation via Gibbs Sampler and Related MCMC Methods. *Journal of the Royal Statistical Society,*B ,55, p.3-24, 1993.

14 TANNER, M.; WONG,W The calculations of Posterior Distributions by Date Augmentation. *Journal of the American Statistical Association,*82, p.528-550, 1987.

15 TITTERINGTON, D.M.; SMITH,A.F.M.;MAKOV,U.V *Satistical Analysis of Finite Mixture Distributions.* New York: John Wiley,1985.

NOTAS DO ICMC

SÉRIE ESTATÍSTICA

061/2000 ANDRADE, M.G.; MEIRA, S.A.; FRAGOSO, M.D.; CARNEIRO, A.A.F.M. – A bayesian approach to the stochastic flood control problem.

060/2000 ACHCAR, J.A.; JANEIRO, V. – A bayesian analysis for corralated binary data in the presence of covariates.

059/2000 MAZUCHELI, J.; ACHCAR, J.A.; KASS, R.E. – Regression models for lifetime data with mixture of normal distributions.

058/99 ACHCAR, J.A.; FORTULAN, V.C. – Meta analysis: a bayesian approach.

057/99 OLIVEIRA, S.C.; ACHCAR, J.A. - Confiabilidade de redes: um enfoque bayesiano.

056/98 RODRIGUES, J.; CHAVES, J.S. – A note on bayesian exponential regression model with censored data.

055/98 RODRIGUES, J.; SILVEIRA, V.D.R. – Bayesian computation for dichotomous variables with classification errors.

054/98 ACHCAR, J.A.; FORTULAN, V.C. – Relação entre o uso de hormônio e câncer em mulheres: um aplicação de meta-análise sob um enfoque bayesiano.

053/98 CID, J.E.R.; ACHCAR, J..A. – Bayesian inference for nonhomogeneous poisson processes in software reliability models assuming nonmonotonic intensity functions.

052/98 ANDRADE FILHO, M.G; MIZOI, M.F. – Aplicação de MCMC na estimação de máxima verossimilhança para processos AR(p) e MA(q).