Revista de Estatística
5ª Página
1º Quadrimestre de 1998

CLASSIFICATION OF VARIABLES IN DISCRIMINANT ANALYSIS: A NEW APPROACH

Autores:
Paulo Gomes
e
Adelaide Figueiredo

VOLUME 1
1º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
7ª PÁGINA
1º QUADRIMESTRE DE 1998

CLASSIFICATION OF VARIABLES IN DISCRIMINANT ANALYSIS: A NEW APPROACH

CLASSIFICAÇÃO DE VARIÁVEIS NA ANÁLISE DISCRIMINANTE: UMA NOVA ABORDAGEM

Autores: Paulo Gomes
- Professor Associado, Universidade Católica Portuguesa
- Director Regional do Instituto Nacional de Estatística

e

Adelaide Figueiredo
- Assistente da Faculdade de Economia da Universidade do Porto

ABSTRACT:

- In this report we develop an approach analogous to the Discriminant Analysis to classify new variables into previously defined groups of variables and give the misclassification probabilities estimates.

KEY WORDS:

- *Discriminant Analysis; Classification; Variables Selection*

RESUMO:

- Neste trabalho procurou desenvolver-se uma abordagem análoga à Análise Discriminante para classificar novas variáveis em grupos de variáveis previamente definidos e obter estimativas das probabilidades de erro de má classificação.

PALAVRAS-CHAVE:

- *Análise Discriminante; Classificação; Selecção de Variáveis*

1. INTRODUCTION

One of the most important questions of Discriminant Analysis tries to solve is the affectation of new individuals into previously defined groups of individuals. We consider the dual problem of the affectation of new variables into previously defined groups of variables. While in Discriminant Analysis, we usually associate the Multivariate Normal distribution to each group of individuals for defining the affectation rules, in our approach we associate the Bingham distribution on the sphere to each group of variables for defining the affectation rules. In our study we suppose that the groups of variables were been previously identified. The identification of those groups can be done using the *k-means method* (Gomes, 1987) or the *E.M. algorithm* of Estimation-Maximization type (Gomes e Figueiredo, 1995). Additionally it´s important to give the estimate of the misclassification probability.

In our approach, we suppose that the n individuals are fixed and the p variables, previously normalised, are randomly selected from a population of variables. Then, we associate to each variable a $1 \times n$ random vector $\left(X_1^j, X_2^j, \cdots, X_n^j\right)$ where X_i^j is a random variable that represents the value of the j^{th} variable for the i^{th} individual. In the classical approach the p variables are fixed and we select randomly the individuals and associate to each individual a $1 \times p$ random vector $\left(X_i^1, X_i^2, \cdots, X_i^p\right)$ where X_i^j is a random variable that represents the value of the j^{th} variable for the i^{th} individual.

In section 2, we adapt to our approach the classification rules used in Discriminant Analysis, and section 3 applies the rules to the case of a Bingham distribution on the sphere. Some methods to the calculation of the estimates of the misclassification probabilities are described in section 4 and an example is given in section 5. Finally, section 6 contains the conclusion.

2. CLASSIFICATION RULES

Consider our data with n individuals and p variables and suppose that the p variables are divided in two groups. We associate to the groups of variables, two subpopulations G_1 and G_2. The variables taken from subpopulation G_1 have the density function on the sphere $f_1(\mathbf{x})$ and the variables from subpopulation G_2 have the density function on the sphere $f_2(\mathbf{x})$.

Let S_{n-1} be the surface of the sphere:

$$S_{n-1} = \left\{\mathbf{x} \in \Re^n : \|\mathbf{x}\| = 1\right\}$$

We divide S_{n-1} into mutually exclusive and exhaustive regions R_1 and R_2. Variables falling in region R_1 are classified into subpopulation 1 and those falling in region R_2 are classified into subpopulation 2. The *likelihood-ratio rule* is the following:

Assign a variable $\mathbf{x}$ to G_1 if

$$f_1(\mathbf{x}) \geq f_2(\mathbf{x}) \qquad (1)$$

and assign $\mathbf{x}$ to G_2, otherwise.

The classification regions R_1 and R_2 are defined by

$$R_1 = \left\{ \mathbf{x} \in S_{n-1} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \right\}$$

$$R_2 = \left\{ \mathbf{x} \in S_{n-1} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1 \right\}$$

We now mention a rule based on the Statistical Decision Theory where the misclassification costs and the prior probabilities are considered.

Let π_i be the probability of a given variable belonging to group i, denoted by prior probability, $i = 1,2$ and $C(i|j)$ the cost of misclassifying a variable from group j into group i, $i,j = 1,2$. The probability of misclassification is given by

$$P(i|j) = P\left(\mathbf{X} \in R_i \middle| \mathbf{X} \in G_j\right) = \int_{R_i} f_j(\mathbf{x}) d\mathbf{x}$$

Then, $\pi_i . P(j|i)$ is the probability of choosing a variable from G_i and wrongly classifying it into G_j, $i, j, = 1,2$.

Denoting by C the total cost of misclassification, we have $E(C) = C(2|1).\pi_1 . P(2|1) + C(1|2).\pi_2 . P(1|2)$.

The rule used to minimize *E (C)*, known as *Bayes rule*, is the following:

Assign a variable **x** to G_1 if

$$C(2|1).\pi_1 . f_1(x) \geq C(1|2).\pi_2 . f_2(x)$$

and assign **x** to G_2 , otherwise.

The classification regions R_1 and R_2 are defined by

$$R_1 = \left\{ \mathbf{x} \in S_{n-1} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{C(1|2).\pi_2}{C(2|1).\pi_1} \right\}$$

$$R_2 = \left\{ \mathbf{x} \in S_{n-1} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{C(1|2).\pi_2}{C(2|1).\pi_1} \right\}$$

If we have equal misclassification costs and equal prior probabilities, the Bayes rule and the likelihood-ratio rule will be the same.

Suppose now, we have k subpopulations G_1 , G_2 ,..., G_k with density functions on the sphere $f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots f_k(\mathbf{x})$, respectively. We divide S_{n-1} into R_1 , R_2 ,..., R_k mutually exclusive and exhaustive regions. Any given variable is classified into the subpopulation in which region the variable falls. The likelihood-ratio rule is the following:

Assign a variable **x** to G_j if

$$f_j(x) = \max_{i=1,\cdots k} f_i(\mathbf{x})$$

We now consider the Bayes rule.

Let π_i be the prior probability of a given variable belonging to G_i and let $C(j|i)$ be the cost of misclassifying a variable from G_i into G_j . The misclassification probability is given by $P(j|i) = P(X \in R_j | X \in G_i) = \int_{R_j} f_i(\mathbf{x})d\mathbf{x}$

The total cost of misclassifying a variable belonging to G_i is given by $\sum_{\substack{j=1 \\ j\neq i}}^{k} P(j|i).C(j|i).$ and the expected total cost of misclassifying variables belonging to G_i will be $\pi_i\left[\sum_{\substack{j=1 \\ j\neq i}}^{k} P(j|i).C(j|i).\right]$. The resulting total expected cost of misclassification, for all subpopulations is defined by

$$E(C) = \sum_{i=1}^{k} \pi_i \left(\sum_{\substack{j=1 \\ j\neq i}}^{k} P(j|i).C(j|i).\right)$$

The rule for minimizing *E (C)*, Bayes rule, is the following:

Assign a variable **x** to G_j if

$$\sum_{\substack{i=1 \\ i\neq j}}^{k} \pi_i . f_i(\mathbf{x}).C(j|i) < \sum_{\substack{i=1 \\ i\neq l}}^{k} \pi_i . f_i(\mathbf{x}).C(l|i) \quad l = 1,\cdots,k \quad ,l \neq j \qquad (2)$$

2.1 POSTERIOR PROBABILITIES

In addition to the affectation group of each variable, we will give more information about each variable if we determine its posterior probability.

The posterior probability of a variable **X** belonging to population i is

$$P(X \in G_i | X = \mathbf{x}) = \frac{f_i(\mathbf{x}).\pi_i}{\sum_{i=1}^{k} f_i(\mathbf{x}).\pi_i} \quad ,i = 1,\cdots,k \qquad (3)$$

where $\pi_i = P(X \in G_i)$ is the prior probability of population i.

Based on the posterior probabilities, we can affect a variable **x** to the population G_i , for which the posterior probability (3) or $\pi_i . f_i(\mathbf{x})$ is the largest. It is equivalent to using the Bayes rule with equal misclassification costs.

3. APPLICATION OF THE CLASSIFICATION RULES

We use a particular case of the Bingham distribution on the sphere whose density function is given by

$$f(\mathbf{x}) = \left\{ {}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \xi\right) \right\}^{-1} . \exp\left(\xi . {}^t\mathbf{u}.\mathbf{x}.{}^t\mathbf{x}.\mathbf{u}\right) \quad \mathbf{x} \in S_{n-1}$$

where ${}_1F_1$ is a confluent hipergeometrical function defined by

$${}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \xi\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right).\Gamma\left(\frac{n-1}{2}\right)} \int_0^1 e^{\xi t} t^{-\frac{1}{2}} (1-t)^{\frac{n-3}{2}} dt$$

This probability distribution denoted by $B_n(\mathbf{u}, \xi)$, has two parameters:

- $\mathbf{u}$ is a directional parameter $({}^t\mathbf{uu} = 1)$
- $\xi (> 0)$ - concentration parameter around the directional parameter

If we associate axes to the variables the Bingham distribution will be appropriate for modelling the group of variables because its density function satisfies the antipodal symmetric property $f(\mathbf{x}) = f(-\mathbf{x})$.

The maximum likelihood estimators of the parameters $\mathbf{u}$ and ξ from the Bingham distribution $B_n(\mathbf{u}, \xi)$, based on a random sample of p variables $X = \left(\mathbf{x}_1 \middle| \mathbf{x}_2 \middle| \cdots \middle| \mathbf{x}_p\right)$, were obtained by (Gomes, 1987) and are following:

- $\hat{\mathbf{u}}$ is the eigenvector associated to the largest eigenvalue of the matrix $X^t.X$, that is

$$X.{}^tX.\hat{\mathbf{u}} = \omega.\hat{\mathbf{u}} \qquad (4)$$

where w is the largest eigenvalue of $X.{}^tX$. Then, the maximum likelihood estimator for $\mathbf{u}$ and the first principal component associated to the group of variables are coincident.

- $\hat{\xi}$ is the solution of the equation

$$Y(\xi) = \frac{\omega}{p}$$

where

$$Y(\xi) = \frac{d}{d\xi} \ln\left[{}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \xi\right) \right]$$

3.1 TWO GROUPS

We refer only the Bayes rule because the rule based on the likelihood ratio is a particular case.

Suppose that the population 1 is Bingham $B_n(\mathbf{u}_1, \xi_1)$ and that the population 2 is Bingham $B_n(\mathbf{u}_2, \xi_2)$.

Let

$$U = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\left[{}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \xi_1\right)\right]^{-1} . e^{\xi_1 \left({}^t u_1 . \mathbf{x}\right)^2}}{\left[{}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \xi_2\right)\right]^{-1} . e^{\xi_2 \left({}^t u_2 . \mathbf{x}\right)^2}}$$

When the parameters $\mathbf{u}_1, \mathbf{u}_2, \xi_1$ and ξ_2 are unknown, they may be replaced by their maximum likelihood estimates.

The Bayes rule is:

Assign **x** to population 1 if

$$\ln U - \ln \frac{C(1|2).\pi_2}{C(2|1).\pi_1} \geq 0$$

or

$${}^t\mathbf{x}\hat{A}\mathbf{x} + \ln \frac{{}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \hat{\xi}_2\right)}{{}_1F_1\left(\frac{1}{2}, \frac{n}{2}, \hat{\xi}_1\right)} - \ln \frac{C(1|2).\pi_2}{C(2|1).\pi_1} \geq 0 \quad (5)$$

where

$$\hat{A} = \left(\sqrt{\hat{\xi}_1}.\hat{\mathbf{u}}_1 + \sqrt{\hat{\xi}_2}.\hat{\mathbf{u}}_2\right).^t\left(\sqrt{\hat{\xi}_1}.\hat{\mathbf{u}}_1 - \sqrt{\hat{\xi}_2}.\hat{\mathbf{u}}_2\right)$$

and assign **x** to population 2, otherwise.

Let the discriminant function be

$$W_{12} = {}^t\mathbf{x}\hat{A}\mathbf{x} + \ln\frac{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_2\right)}{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_1\right)} - \ln\frac{C(1|2).\pi_2}{C(2|1).\pi_1} \qquad (6)$$

the discriminant regions R_1 and R_2 are given by

$$R_1 = \{x: W_{12} \geq 0\}$$
$$R_2 = \{x: W_{12} < 0\}$$

3.2 MORE THAN TWO GROUPS

Suppose the k subpopulations are Bingham $B_n(\mathbf{u}_i, \xi_i)$, $i = 1,\cdots k$, whose density function is given by

$$f_i(\mathbf{x}) = \left\{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\xi_i\right)\right\}^{-1}.e^{\xi_i({}^t u_i\mathbf{x})^2} \qquad \mathbf{x} \in S_{n-1}$$

The Bayes rule is

Classify **x** as a variable from population j if

$$\sum_{\substack{i=1\\ i\neq j}}^{k} \pi_i.f_i(\mathbf{x}).C(j|i) = \min_{l=1,\cdots k}\sum_{\substack{i=1\\ i\neq l}}^{k} \pi_i.f_i(\mathbf{x}).C(l|i) \qquad (7)$$

or

$${}^t\mathbf{x}\hat{A}\mathbf{x} + \ln\frac{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_i\right)}{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_j\right)} - \ln\frac{C(j|i).\pi_i}{C(i|j).\pi_j} \geq 0 \quad i = 1,\cdots,k, \quad i \neq j \qquad (8)$$

where

$$\hat{A} = \left(\sqrt{\hat{\xi}_j}.\hat{\mathbf{u}}_j + \sqrt{\hat{\xi}_i}.\hat{\mathbf{u}}_i\right).{}^t\left(\sqrt{\hat{\xi}_j}.\hat{\mathbf{u}}_j - \sqrt{\hat{\xi}_i}.\hat{\mathbf{u}}_i\right)$$

and $\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\xi}_i, \hat{\xi}_j$ are the maximum likelihood estimators of the parameters.

Let

$$W_{ji} = {}^t\mathbf{x}\hat{A}\mathbf{x} + \ln\frac{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}\right)}{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_j\right)} - \ln\frac{C(j|i).\pi_i}{C(i|j).\pi_j} \qquad (9)$$

be the discriminant function.

The region R_j is given by

$$R_j = \left\{\mathbf{x}: W_{ji} \geq 0 \quad i = 1,\cdots,k, \quad i \neq j\right\}$$

In case of three groups, the discriminant functions are given by

$$W_{12} = \hat{\xi}_1.{}^t\mathbf{x}.\hat{\mathbf{u}}_1.{}^t\hat{\mathbf{u}}_1.\mathbf{x} - \hat{\xi}_2.{}^t\mathbf{x}.\hat{\mathbf{u}}_2.{}^t\hat{\mathbf{u}}_2.\mathbf{x} + \ln\frac{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_2\right)}{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_1\right)} - \ln\frac{C(1|2).\pi_2}{C(2|1).\pi_1}$$

$$W_{13} = \hat{\xi}_1.{}^t\mathbf{x}.\hat{\mathbf{u}}_1.{}^t\hat{\mathbf{u}}_1.\mathbf{x} - \hat{\xi}_3.{}^t\mathbf{x}.\hat{\mathbf{u}}_3.{}^t\hat{\mathbf{u}}_3.\mathbf{x} + \ln\frac{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_3\right)}{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_1\right)} - \ln\frac{C(1|3).\pi_3}{C(3|1).\pi_1}$$

$$W_{23} = \hat{\xi}_2.{}^t\mathbf{x}.\hat{\mathbf{u}}_2.{}^t\hat{\mathbf{u}}_2.\mathbf{x} - \hat{\xi}_3.{}^t\mathbf{x}.\hat{\mathbf{u}}_3.{}^t\hat{\mathbf{u}}_3.\mathbf{x} + \ln\frac{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_3\right)}{{}_1F_1\left(\frac{1}{2},\frac{n}{2},\hat{\xi}_2\right)} - \ln\frac{C(2|3).\pi_3}{C(3|2).\pi_2}$$

As the condition

$$W_{23} = W_{13} - W_{12} + \ln\frac{C(1|3).\pi_3}{C(3|1).\pi_1} - \ln\frac{C(1|2).\pi_2}{C(2|1).\pi_1} - \ln\frac{C(2|3).\pi_3}{C(3|2).\pi_2}$$

holds, it is enough to use only the two functions W_{13} and W_{12} for defining the discriminant regions R_1, R_2 and R_3.

$$R_1 = \left\{\mathbf{x}: W_{12} \geq 0, \quad W_{13} \geq 0\right\}$$

$$R_2 = \left\{\mathbf{x}: W_{12} \leq 0, \quad W_{13} \geq W_{12} - \ln\frac{C(1|3).\pi_3}{C(3|1).\pi_1} + \ln\frac{C(1|2).\pi_2}{C(2|1).\pi_1} + \ln\frac{C(2|3).\pi_3}{C(3|2).\pi_2}\right\}$$

$$R_3 = \left\{\mathbf{x}: W_{13} \leq 0, \quad W_{13} \leq W_{12} - \ln\frac{C(1|3).\pi_3}{C(3|1).\pi_1} + \ln\frac{C(1|2).\pi_2}{C(2|1).\pi_1} + \ln\frac{C(2|3).\pi_3}{C(3|2).\pi_2}\right\}$$

4. MISCLASSIFICATION PROBABILITIES

In the case of two groups, the misclassification probabilities are

$$P(2|1) = P \text{ (misclassifying a variable from } G_1 \text{ into } G_2) = P\left(W_{12} < 0 \middle| \mathbf{x} \in G_1\right)$$

$$P(1|2) = P \text{ (misclassifying a variable from } G_2 \text{ into } G_1) = P\left(W_{12} \geq 0 \middle| \mathbf{x} \in G_2\right)$$

As the distribution for the random variable W_{12} if $x \in G_1$ or if $x \in G_2$ is unknown, we can not calculate these probabilities, but we can estimate them.

4.1 ESTIMATES OBTAINED BY SIMULATION

If we consider two subpopulations of Bingham, $B_n\left(\hat{\mathbf{u}}_1, \hat{\xi}_1\right)$ and $B_n\left(\hat{\mathbf{u}}_2, \hat{\xi}_2\right)$, we can simulate an adequate number of variables from each subpopulations. For each simulated variable, we calculated the value for the random variable W_{12}. An estimate of $P(2|1)$ is the proportion of the variables from population 1 for which W_{12} <0 and an estimate of $P(1|2)$ is the proportion of variables from population 2 for which $W_{12} \geq 0$.

4.2 ESTIMATES OF DISCRIMINANT ANALYSIS

In Discriminant Analysis there are three probabilities of correct classification (i.e., three population hit rates):

- *optimal hit rate* obtained when a classification rule based on known parameters is applied to the population.
- *actual hit rate* obtained by applying a rule based on a particular sample to future samples.
- *expected actual hit rate* is the expected proportion of correct classifications over all possible samples.

The actual hit rate is the most common and can be estimated by using an internal analysis or an external analysis (i.e., the Holdout method or the Leave-One-Out Method) or the Bootstrap technique. These methods will be described next.

4.2.1 INTERNAL ANALYSIS

We use samples of each subpopulation to estimate the unknown parameters of the subpopulations, and to determine a rule, then, based on the obtained rule, we classify the units of the same samples. The obtained rates are called *apparent or resubstitution* hit rates.

As the apparent hit rate yields a positively biased estimate, we use an external analysis to validate the classification rule.

4.2.2 EXTERNAL ANALYSIS

Whereas in the internal analysis the units classified are the units used for defining the rule in the external analysis, the rule is determined from one set of units and then used to classify another set of units.

Holdout method

We divide the sample (of each subpopulation) in two subsamples: a training sample and a test sample. We estimate the unknown parameters of the subpopulations based on the training sample and, therefore, obtain the classification rule. Then, with the obtained rule we classify the units of the test sample. The proportion of the test sample units that are correctly classified is a hit rate estimate. This method requires the use of large samples.

Leave-One-Out method (Lachenbruch, P.A. and Mickey, 1968)

We delete one unit from one of the subpopulations samples and, based on the remaining units of that sample, we estimate the unknown parameters of that subpopulation. We also estimate the unknown parameters of the other subpopulations based on the respective samples. Therefore, we obtain a rule that we use for classifying the deleted unit into one of the groups. Afterwards, we delete another unit of the sample and we determine the rule used for classifying the deleted unit. We repeat the process until all units of each sample have been deleted. The proportions of deleted units correctly classified are the hit-rate estimates.

4.2.3 BOOTSTRAP TECHNIQUE (EFRON, 1987)

After having estimated the unknown parameters based on the training samples and having determined the rule, we generate samples by choosing units at random and with replacement from the sample composed by the test samples. For each generated sample, called *bootstrap sample*, we classify its units and calculate the hit rate because we know to which group belongs each variable.

As the hit-rate empirical bootstrap distribution is an approximation to the hit-rate distribution, an estimate of the true hit rate is the average of the hit rates obtained with the bootstrap samples.

In our approach, we use the previous estimates of the misclassification probabilities.

5. EXAMPLE

We considered the data based on the statistic of crime for England and Wales during the period 1950-1963. See appendix.

Using the method of Principal Components, (Ahmad, 1967), has shown that the yearly variation in the number of crimes could be explained by a small number of unrelated factors. They are, mainly, the 1st component, associated with the population growth, and the 2nd component that seems likely to reflect changes in recording practice by the police over the period.

Supposing that the group of variables is composed by two subgroups of variables, each one from a Bingham population and identifying a mixture of two Bingham distributions on the sphere by the *k-means method* or the *E. M. algorithm*, we obtained the following two groups of variables:

Group I - 1,3,6,10,13,17

Group II - 2,4,5,7,8,9,11,12,14,15,16,18

being the variables of group I more associated with changes in the recording of crimes and the variables of group II more associated with the population growth.

After having identified the two groups of variables, we used the likelihood-ratio rule for classifying variables into one of the groups.

Firstly, we determined the estimates of the misclassification probabilities, by simulation. We simulated 1000 variables from each of the Bingham populations $B_{14}\left(\hat{\mathbf{u}}_1, \hat{\xi}_1\right)$ and $B_{14}\left(\hat{\mathbf{u}}_2, \hat{\xi}_2\right)$ where $\hat{\mathbf{u}}_1$ and $\hat{\xi}_1$ are the maximum likelihood estimators based on the sample of dimension 6 from population 1 and $\hat{\mathrm{u}}_2$ and $\hat{\xi}_2$ the maximum likelihood estimators based on the sample of dimension 12 from population 2. We obtained the estimates:

$$\hat{P}(2|1) = 3.9\%$$
$$\hat{P}(1|2) = 5.2\%$$

Secondly we classified the variables of each sample. The results are shown in the following table:

		Origin group	
		1	2
Affectation group	1	5	0
	2	1	12
		6	12

i.e., the apparent estimates of misclassification probabilities where:

$$\hat{P}(2|1) = \frac{1}{6}$$
$$\hat{P}(1|2) = 0$$

and the global hit rate was 94,4%

These estimates gives us just an idea of how the groups of variables were separated. Therefore, we used the Holdout method for validating the rule.

In the Holdout method, we considered about 30% of the sample of each group for the test sample and about 70% of the sample for the training sample. So, for the test sample, we chose randomly 2 variables from group I, 4 variables from group II and the other 12 variables constitute the training sample which was used for defining

the rule. We considered three test samples to avoid biased sampling. The samples considered are described in the following table:

	Group	Test samples	Training samples
1st Case	I	13,17	1,3,6,10
	II	11,12,14,15	2,4,5,7,8,9,16,18
2nd Case	I	6,13	1,3,10,17
	II	4,8,9,11	2,5,7,12,14,15,16,18
3rd Case	I	1,10	3,6,13,17
	II	2,7,14,15	4,5,8,9,11,12,16,18

The misclassification probabilities estimates in each case, are:

	$\hat{P}(2\|1)$	$\hat{P}(1\|2)$
1st Case	0	0
2nd Case	0.5	0
3rd Case	0.5	0

We can take the average of these estimates of probabilities as estimates of the misclassification probabilities:

$$\hat{P}(2|1) = 33.3\%$$
$$\hat{P}(1|2) = 0$$

Finally we used also the method Leave-one-out, where we obtained the following misclassification probabilities estimates:

$$\hat{P}(2|1) = 16.7\%$$
$$\hat{P}(1|2) = 0$$

6. CONCLUSION

Our approach is concerned with variables selection under a multivariate statistical analysis, in a data science classification context, where we established, according to a probabilistic model defined in the R^n sphere, a decision rule which enables us to affect a new variable into two or more groups of variables previously defined through the k-means method or E.M. algorithm.

The misclassification error problem was studied by a simulation method, afterwards by using apparent estimates and, finally, based on Holdout and on Leave-one-out methods. The second approach tends to underestimate the misclassification probabilities. Though, using the Leave-one-out method seems to confirm the results

obtained by that second approach. The small number of variables included in the samples suggests the attribution of particular relevance to the estimates of misclassification probabilities provided by simulation.

BIBLIOGRAPHY

AHMAD, (1967) "An analysis of crimes by the method of Principal Components", *Applied Statistics*, pp.17-39.

EFRON, (1987) "Better Bootstrap Confidence Intervals", *Journal of the American Statistical Society*, 82 (March), 171-185.

GOMES, (1987) "Distribution de Bingham sur la n-sphère: Une Nouvelle Approche de l'Analyse Factorielle", *Thèse d'État* , Université de Montpellier.

Gomes e Figueiredo, (1995) "Identificação de uma mistura de leis de Bingham através dum algoritmo tipo *E.M.*", *Actas do III Congresso da Sociedade Portuguesa de Estatística.*

HUBERTY, C., (1994) "Applied Discriminant Analysis" *Wiley Series in Probability and Mathematical Statistics.*

LACHENBRUCH, P.A. and MICKEY, (1986) "Estimation of error rates in discriminant analysis" *Technometrics*, vol.10, pp.1-11.

MARDIA, K. (1972) "Statistics of directional data" *Academic Press.*

WATSON, (1983) "Statistics on spheres". *J. Wiley.*

A APPENDIX

A.1 VARIABLES

1- Homicide: Murder, attempted murder, manslaughter, infanticide

2- Woundings: Felonious, Malicious, Assault

3- Homosexual offences: Buggery and attempts at Indecency between males

4- Heterosexual offences: Rape, indecent assault, unlawful intercourse, incest

5- Breaking and entering: Sacrilege, burglary, housebreaking, shopbreaking, etc.

6- Robbery

7- Larceny: Embezzlement, aggravated larceny, etc

8- Frauds and false pretences

9- Receiving stolen goods

10- Malicious injuries to property

11- Forgery, etc

12- Blackmail

13- Assault

14- Malicious damage

15- Revenue laws

16- Intoxicating laws

17- Indecent exposure

18- Taking motor vehicle without consent

A.2 DATA

ind/var	1	2	3	4	5	6	7	8	9
1	529	5.258	4.416	8.178	92.839	1.021	301.078	25.333	7.586
2	455	5.619	4.876	9.223	95.946	800	355.407	27.216	9.716
3	555	5.98	5.443	9.026	97.941	1.002	341.512	27.051	9.188
4	456	6.187	5.68	10.107	88.607	980	308.578	27.763	7.786
5	487	6.586	6.357	9.279	75.888	812	285.199	26.267	6.468
6	448	7.076	6.644	9.953	74.907	823	295.035	22.966	7.016
7	477	8.433	6.196	10.505	85.768	965	323.561	23.029	7.215
8	491	9.774	6.327	11.9	105.042	1.194	360.985	26.235	8.619
9	453	10.945	5.471	11.823	131.132	1.692	409.388	29.415	10.002
10	434	12.707	5.732	13.864	133.962	1.9	445.888	34.061	10.254
11	492	14.391	5.24	14.304	151.378	2.014	489.258	36.049	11.696
12	459	16.197	5.605	14.376	164.806	2.349	531.430	39.651	13.777
13	504	16.43	4.866	14.788	192.302	2.517	588.566	44.138	15.783
14	510	18.655	5.435	14.722	219.138	2.483	635.627	45.923	17.777

ind/var	10	11	12	13	14	15	16	17	18
1	4.158	3.79	118	20.844	9.447	24.616	49.007	2.786	3.126
2	4.993	3.378	74	19.963	10.359	21.122	55.229	2.739	4.595
3	5.003	4.173	120	19.056	9.108	23.339	55.635	2.598	4.145
4	5.309	4.649	108	17.772	9.278	19.919	55.688	2.639	4.551
5	5.251	4.903	104	17.379	9.176	20.585	57.011	2.587	4.343
6	2.184	4.086	92	17.329	9.46	19.197	57.118	2.607	4.836
7	2.559	4.04	119	16.677	10.997	19.064	63.289	2.311	5.932
8	2.965	4.689	121	17.539	12.817	19.432	71.014	2.31	7.148
9	3.607	5.376	164	17.344	14.289	24.543	69.864	2.371	9.772
10	4.083	5.5987	160	18.047	14.118	26.853	69.751	2.544	11.211
11	4.802	6.59	241	18.801	15.866	31.266	74.336	2.719	12.519
12	5.606	6.924	205	18.525	16.399	29.922	81.753	2.82	13.050
13	6.256	7.816	250	16.449	16.852	34.915	89.709	2.614	14.141
14	6.935	8.634	257	15.918	17.003	40.434	89.149	2.777	22.896