Method for detection of mixtures of normal distributions with application to vine varieties

Amílcar Oliveira[1], Teresa Oliveira[2]

[1]Universidade Aberta, Department of Sciences and Technology
Rua Fernão Lopes 9 2Dir, 1000 132 Lisbon Portugal, e-mail: aoliveira@univ-ab.pt

[2]Universidade Aberta, Department of Sciences and Technology
Rua Fernão Lopes 9 2Dir, 1000 132 Lisbon Portugal, e-mail: toliveira@univ-ab.pt

Abstract. In this work we trait the problem of mixtures of normal distributions and methods for estimating the number of components as well as the parameters in a mixture. Also, we present a practical method for the detection of normal finite mixture distribution and respective model validation. Finally, we apply the exposed procedure to a sample of old grape-vine castes.

Keywords: Normal distribution, mixtures, grape-vine caste.

1 Introduction

Finite mixtures of distributions, and in particular mixtures of normal distributions, have been extensively used to model a wide variety of important practical situations, in which data can be considered from two or more populations mixed in varying populations. It is therefore evident interest in this subject attending the vast applications that have been developed by statisticians. We will emphasize some topics that have been successfully addressed in this area, which include among others the problem of identification of outliers, (Atkin & Tunnicliffe, 1980) (Wilson, 1980) or (Beckman & Cook, 1983), latent class models (Goodman, 1974), classification analysis (Symons, 1981) (Celeux, 1986) or (Bozdogan, 1992), investigating the robustness of certain statistics such as correlation coefficient sample studied (Srivastava & Lee 1984).

In our work there will be some introductory remarks in the context of finite mixtures in order to create the enabling environment for a better understanding of this subject. We will continue with a statistical approach to key issues in the context of mixtures which will focus on the main methods of estimating parameters of a mixture and one of the most used algorithms in the identification of a finite mixture of distributions, i.e., the so-called EM algorithm (Dempster, et al., 1977).

In: M. Salampasis, A. Matopoulos (eds.): Proceedings of the International Conference on Information and Communication Technologies

for Sustainable Agri-production and Environment (HAICTA 2011), Skiathos, 8-11 September, 2011.

2 Estimating the number of components in a mixture

The problem of statistical analysis of finite mixtures can be divided into the following phases:

i) Verification of identifiability of the mixture;
ii) Estimation of the number of components / parameters estimation,
iii) Testing the number of components
iv) Validation of the model

With a finite mixture density, *f(x, r)* function is identifiable if and only if

$$\sum_{j=1}^{k} p_j f\left(x \middle| \theta_j\right) = \sum_{j=1}^{k^*} p_i^* f\left(x \middle| \theta_i^*\right) \Rightarrow k = k^* \wedge \left(\forall_{j=1,\ldots k}, \exists_{i=1,\ldots k^*} : p_j = p_i^* \wedge \theta_j = \theta_i^*\right) \quad \textbf{(1)}$$

That is to say that the mixture is identifiable if it admits only a single decomposition. (Teicher, 1963) deduced the necessary and sufficient conditions for identifiability and proved that the mixtures of normal distributions are identifiable.

3 Methods for the estimation of mixture parameters

The problem of parameter estimation in mixtures, in the case of normal distributions, is one of the oldest problems in the statistical literature. It was first introduced by (Pearson, 1894) in an article "Contribution to the theory of evolution mathematical" and subsequently developed by (Quandt & Ramsey, 1978). It is still an open problem which attracts strong attention.

Although at present the study of mixtures takes place in several areas by applying other methods, such as the method of maximum likelihood, the original method, the method of moments is still considered one of the best approximation methods in separating mixtures of normal distributions. It is useful even in the generation of initial estimates for the iterative resolution of maximum likelihood equations.

In order to make the analysis of mixtures of distributions a computational problem more accessible, in the decades of the forties and fifties it was fostered the development of a high number of graphical techniques. A first step consisted in the detection of turning points of the curves (Harding, 1948) and (Cassie, 1954), making this method a somewhat subjective process. Later, more rigorous techniques have been suggested to determine the inflection points (Fowlkes, 1979) and (Bhattacharya, 1967); this author also suggests several methods to determine the proportions of the mixtures.

These graphical techniques are not only to give a first estimation of parameters, they can be quite useful in an initial examination of data, since they have the advantage of running without a prior knowledge of the number of components of the mixture. They can play a role as an indicator of the number of components, since this information is needed before applying any of the other methods described below.

3.1 Method of Moments

This method is used on obtaining and solving a system of equations, often of the nonlinear type. The equations are obtained from the empirical equality of every moment $\left(M_r\right)$ and their theoretical moment $\left(m_r\right)$,

$$M_r = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^r, \quad r = 1,\ldots,m \qquad \textbf{(2)}$$

and

$$m_r = \int \left(x-\mu\right)^r f\left(x\right) dx, \quad r = 1,\ldots,m \qquad \textbf{(3)}$$

Where $\bar{x}$ represents the mean of a sample obtained from a population with probability density function $f\left(x\right)$, μ is the mean value of random variable X, with the same probability function, m the number of moments needed to estimate all parameters.
Let for each value of r obtaining an equation:

$$\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^r = \int \left(x-\mu\right)^r f\left(x\right) dx \qquad \textbf{(4)}$$

Note that with the currently available computational means this method becomes very advantageous. The disadvantage lies in the fact that it is not applicable to mixtures of distributions with a large number of components, as well as to the multidimensional case.

3.2 Method of Quandt and Ramsey

This method proposed by (Quandt & Ramsey, 1978) is used in mixtures of two univariate components and makes use of the moment generating function $E\left(e^{tx}\right)$.
The estimate for this function is given by:

$$\hat{E}\left(e^{tx}\right) = \frac{1}{n}\sum_{j=1}^{n} e^{tx_j} \qquad \textbf{(5)}$$

The method minimizes the sum of squared deviations between the empirical moment generating function and the theoretical moment generating function

$$S^{(t)} = \sum_{j=1}^{n}\left[\hat{E}\left(e^{tx_j}\right) - E\left(e^{tx_j}\right)\right]^2 \qquad \textbf{(6)}$$

where k represents the number of t values in a selected interval (a,b), with $a < 0$ and $b > 0$.

3.3 Method of maximum likelihood

Consider the sample values $x_1,...,x_n$, $x_j \in \Re^m$, $j= 1,...,n$, a mixture of k density functions and consider the log-likelihood function of this sample, represented by

$$L(x|q) = \sum_{j=1}^{n} ln\left[\sum_{j=1}^{n} p_i f(x_j|a_i)\right] \quad \textbf{(7)}$$

where $q = (p_i, a_i), i = 1,...,k$ is the vector of unknown parameters.

By derivation of the function $L(x|q)$ in order to each of these parameters and equating to zero each of the expressions obtained, we have the so-called likelihood equations:

$$\frac{\partial L(x|q)}{\partial q} = 0 \quad \textbf{(8)}$$

The equations obtained are sometimes impossible to solve analytically or very difficult to resolve, so often resort to the use of iterative methods. However, not always the most common processes are able to respond to the scale of problems. In order to overcome these difficulties there is the EM algorithm of (Dempster, et al., 1977), which is the most widely used in solving equations that describe maximum likelihood.

In particular if we have a sample size n and a mixture of two univariate normal components, the log-likelihood function is given by:

$$L(x|q) = \sum_{j=1}^{n} ln\left[p\frac{1}{\sqrt{2\pi\sigma_1}}e^{-\frac{(x_j-\mu_1)^2}{2\sigma_1^2}} + (1-p)\frac{1}{\sqrt{2\pi\sigma_2}}e^{-\frac{(x_j-\mu_2)^2}{2\sigma_2^2}}\right] \quad \textbf{(9)}$$

4 Practical method for detection finite mixture of normal distributions

The method aims on the one hand prove the existence of mixtures of normal distributions and on the other hand achieving the adjustment of an appropriate model. This approach will first check the detailed curves detected in the histogram of simple frequencies, and then estimate the parameters of these curves and the model checking by applying the Kolmogorov-Smirnov test, adapted to a mixture of normal.

4.1 Dissection of the Curves

Consider the original data set $x_1,...,x_n$ of the sample under study, and begin a first step by ordering them in ascending order. We then have the ordered sample $z_1,...,z_n$ with $z_1 \leq ... \leq z_n$. Let n_1, n_2 and n_3 represent the sizes of sub-populations derived from the decomposition of the initial population, with $n_1 + n_2 + n_3 = n$.

Considering the hypothesis of existence of two normal distributions, denote $p = \frac{2n_1}{n}$ and $1 - p = \frac{2n_3}{n}$ where p and $1 - p$ represent the proportions of each one of the curves found on the original curve.

These curves are normal, so they are symmetrical and therefore there is $\frac{n_1 + n_2}{n} = \frac{1}{2}$, or is 50% the size of the sample size n. Let μ_1 and σ_1 be respectively the mean and standard deviation in the distribution curve 1 and μ_2 and σ_2 be respectively the mean and standard deviation in the distribution curve 2.

Thus the probability density function f of weighing the two normal curves is defined by

$$f(x|p,\mu_1,\mu_2,\sigma_1,\sigma_2) = pn(x|\mu_1,\sigma_1) + (1-p)n(x|\mu_2,\sigma_2) \qquad \textbf{(10)}$$

4.2 Estimation of parameters for the general model

4.2.1 Estimates of n_1^*

The value of n_1^* is the one that minimizes $\Delta(n')$ with $\Delta(n') = \left|\bar{x} - \left(p_n'^* \mu_{1,n'}^* + \left(1 - p_n'^*\right)\mu_{2,n'}^*\right)\right|$, where $\bar{x}$ is the sample mean and

$$\begin{cases} \mu^{*}_{1,n'} = \dfrac{z_{n'} + z_{n'+1}}{2} \\ \mu^{*}_{2,n'} = \dfrac{z_{n'+\frac{n}{2}} + z_{n'+\frac{n}{2}+1}}{2} \\ p'^{*}_{n} = \dfrac{2n'}{n} \end{cases} \quad \textbf{(11)}$$

For each n' there is an estimate of the fraction correspondent to each sub-population p'^{*}_{n} and $1 - p'^{*}_{n}$ and of the respective mean values $\mu^{*}_{1,n'}$ and $\mu^{*}_{2,n'}$. Combining these estimates we obtain:

$$\mu'^{*}_{n} = p'^{*}_{n}\mu^{*}_{1,n'} + \left(1 - p'^{*}_{n}\right)\mu^{*}_{2,n'} \quad \textbf{(12)}$$

and the minimization of:

$$\Delta(n') = \left|\bar{x} - \mu'^{*}_{n}\right| \quad \textbf{(13)}$$

4.2.2 Estimates of p , μ_1 and μ_2

Once estimated n_1 we use the correspondent estimates for the fractions and the mean values of sub-populations obtaining:

$$\begin{cases} \mu^{*}_{1} = \dfrac{z_{n_1} + z_{n_1+1}}{2} \\ \mu^{*}_{2} = \dfrac{z_{n_1+\frac{n}{2}} + z_{n_1+\frac{n}{2}+1}}{2} \\ p^{*} = \dfrac{2n_1}{n} \end{cases} \quad \textbf{(14)}$$

the distribution mean value will be given by $\mu = p\mu_1 + (1 - p)\mu_2$.

4.2.3 Estimates of σ_1 and σ_2

To obtain the estimates for σ_1 **and** σ_2 we use:

$$\begin{cases} n_3 = \dfrac{n}{2} - n_1^* \\ \sigma_1^{2*} = \dfrac{1}{n_1^*} \displaystyle\sum_{j=1}^{n_1^*} \left(z_j - \mu_1^*\right)^2 \\ \sigma_2^{2*} = \dfrac{1}{n_3^*} \displaystyle\sum_{j=n_1^*+\frac{n}{2}+1}^{n} \left(z_j - \mu_2^*\right)^2 \end{cases} \qquad \textbf{(15)}$$

As we have seen the reason for equality $n_3^* = \frac{n}{2} - n_1^*$ and is justified by the fact that the tails of two distributions assume negligible values. So, since $n_2 = n_1 + n_3$, and $n_2 = \frac{n}{2}$ it finally comes $n_3 = \frac{n}{2} - n_1$. The term $\left(z_j - \mu_1^*\right)$ reflects the differences between each value of sample 1. σ_1^{2*} and σ_2^{2*} denote the estimated variances for populations 1 and 2.

4.3 Validation of the Model

We will use the Kolmogorov-Smirnov test to check the model. The statistic of this test is the maximum module of the difference between the empirical distribution

$$F_x^* \begin{cases} 0\,;\, x < z_1 \\ \dfrac{i}{n}\,;\, z_i \le x \le z_{i+1} \\ 1\,;\, x > z_n \end{cases} , \quad i = 1,\ldots,n \qquad \textbf{(16)}$$

and the adjusted $p^* N\left(x \middle| \mu_1^*, \sigma_1^*\right) + \left(1 - p^*\right) N\left(x \middle| \mu_2^*, \sigma_2^*\right).$

One of the observed values is then equal to

$$Max\left\{\left|\frac{i}{n} - \left(p^* N\left(z_i \middle| \mu_1^*, \sigma_1^*\right) + \left(1 - p^*\right) N\left(x \middle| \mu_2^*, \sigma_2^*\right)\right)\right|\right\} \quad ; \quad i = 1,\ldots,n$$

We will obtain the values of distributions $N\left(z\middle|\mu_l^*,\sigma_l^*\right)$; $l=1,2$, in z_i points, $i=1,...,n$. From the z_i points, $i=1,...,n$, we calculate

$$\begin{cases} u_i = \dfrac{z_i-\mu_1^*}{\sigma_1^*} \\ v_i = \dfrac{z_i-\mu_2^*}{\sigma_2^*} \end{cases} ; \quad i=1,...,n \tag{17}$$

Now since

$$\begin{cases} N\left(z_i\middle|\mu_1^*,\sigma_1^*\right)=N\left(u_i\middle|0,1\right);\, i=1,...,n \\ N\left(z_i\middle|\mu_2^*,\sigma_2^*\right)=N\left(v_i\middle|0,1\right);\, i=1,...,n \end{cases}$$

The problem is reduced to obtain the values of standardized normal at u_i, $i=1,...,n$ and at u_i, $i=1,...,n$.

We have

$$N\left(w\middle|0,1\right)=\frac{1}{2}+\frac{w}{|w|}\int_0^{|w|}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw=\frac{1}{2}+\frac{w}{|w|}\frac{1}{\sqrt{2\pi}}\sum_{j=0}^{\infty}\frac{(-1)^j}{2^j}\frac{|w|^{2j+1}}{2j+1} \tag{18}$$

So, the problem is simplified since in general it is enough to handle the first 25 items. The above expression is justified since if $w<0$ we have

$$N(w)=\int_{-\infty}^{w}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw=\frac{1}{2}-\int_{w}^{0}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw=\frac{1}{2}-\int_{0}^{|w|}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw \tag{19}$$

and if $w>0$

$$N(w)=\int_{0}^{w}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw=\frac{1}{2}+\int_{0}^{w}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw=\frac{1}{2}+\int_{0}^{|w|}\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dw \tag{20}$$

After obtaining the values for normal distribution we calculate the values:

$$F^*\left(z_j\right)=p^*N\left(u_j\middle|0,1\right)+\left(1-p^*\right)N\left(v_j\middle|0,1\right) \tag{21}$$

And we obtain the statistic

$$T = max\left\{\left|\frac{j}{n} - F^*\left(z_j\right)\right| ; \ j = 1,...,n\right\}$$

This value should be compared with the value in normal distribution table and we concluded about the existence of a mixture of normal distributions if the T value is less than the tabulated. For the previous iterative procedure we decided to simplify the calculations, very time consuming, by designing an algorithm, and we used it in our practical application.

5 Application

First, we will carry out a comprehensive statistical study of available data relating to populations of vine varieties.

We will discuss three types of caste, namely the Trincadeira Preta in years 1988, 1989 and 1990 in the region of Almeirim, Aragon 1987, 1988 and 1989 in the region of Reguengos and Touriga Nacional 1994, 1995 and 1996 in the region of Foz Coa. Regarding the number of observations and replicates of clones for analysis we noted that with regard to the caste Trincadeira Preta there is in each year a total of (271 observations x 5 repetitions) , in Touriga Nacional (197 observations x 5 repetitions) and in the Aragonês (153 observations x 4 repetitions). Note that each observation is respective to a different clone, repeating this clone 4 or 5 times. For a better layout follows the table 1:

Table 1. Varieties per year and repetitions

Varieties	Year	Repetition
Trincadeira Preta	1988,1989,1990	*REP1;REP2;REP3;REP4;REP5*
Aragonês	1987,1988,1989	*REP1;REP2;REP3;REP4*
Touriga Nacional	1994,1995,1996	*REP1;REP2;REP3;REP4;REP5*

For each repetition, we proceeded to calculate the mean, variance, standard deviation, median, sum of sample values, 95%, first quartile, third quartile, range of the sample, inter-quartile range, skewness coefficient, coefficient of flattening and determination of maximum and minimum values of the sample.

Then we obtained the histograms for each of the repetitions, and we tested the normality through the Kolmogorov-Smirnov test.

5.1 Analysis of distributions of repeat genotypes

5.1.1 Testing the normality of distributions

As the normal distribution in one of the most important ones, it is useful at this point proceed to test data normality. To this end we then base our conclusions on the results of a nonparametric test, as mentioned above, the application of the Kolmogorov-Smirnov test.

With the help of statistical software and consulting the Table of Critical Values of the Kolmogorov-Smirnov test for one sample, then we obtained the values KS Observed and KS. Tabulated for each repetition and present in table:

Table 2. Kolmogorov-Smirnov application

			K.S.	K.S.	Tabulated
Varietie	Year	Repetition	Observed	1%	5%
Trincadeira Preta (n=271)	1988	REP1	0.06582	<0.09902	<0.08261
		REP2	0.08012	<0.09902	<0.08261
		REP3	0.06842	<0.09902	<0.08261
		REP4	0.07813	<0.09902	<0.08261
		REP5	0.08156	<0.09902	<0.08261
	1989	REP1	0.04241	<0.09902	<0.08261
		REP2	0.04603	<0.09902	<0.08261
		REP3	0.05565	<0.09902	<0.08261
		REP4	0.06031	<0.09902	<0.08261
		REP5	0.06810	<0.09902	<0.08261
	1990	REP1	0.05569	<0.09902	<0.08261
		REP2	0.06024	<0.09902	<0.08261
		REP3	0.03547	<0.09902	<0.08261
		REP4	0.06007	<0.09902	<0.08261
		REP5	0.03681	<0.09902	<0.08261

Table 3. Kolmogorov-Smirnov application

			K.S.	K.S.	Tabulated
Varietie	Year	Repetition	Observed	1%	5%
Aragonês (n=153)	1987	REP1	0.07008	<0.13178	<0.10995
		REP2	0.07756	<0.13178	<0.10995
		REP3	0.07637	<0.13178	<0.10995
		REP4	0.11603	<0.13178	>***0.10995***
	1988	REP1	0.08873	<0.13178	<0.10995
		REP2	0.04576	<0.13178	<0.10995
		REP3	0.06797	<0.13178	<0.10995
		REP4	0.11283	<0.13178	>***0.10995***
	1989	REP1	0.09338	<0.13178	<0.10995
		REP2	0.10146	<0.13178	<0.10995
		REP3	0.13492	>***0.13178***	>***0.10995***
		REP4	0.12791	<0.13178	>***0.10995***

Table 4. Kolmogorov-Smirnov application

			K.S.	K.S.	Tabulated
Varietie	Year	Repetition	Observed	1%	5%
Touriga Nacional («=197)	1994	REP1	0.07714	<0.11350	<0.09690
		REP2	0.09881	<0.11350	>***0.09690***
		REP3	0.05603	<0.11350	<0.09690
		REP4	0.05166	<0.11350	<0.09690
		REP5	0.06040	<0.11350	<0.09690
	1995	REP1	0.06864	<0.11350	<0.09690
		REP2	0.07562	<0.11350	<0.09690
		REP3	0.06499	<0.11350	<0.09690
		REP4	0.06426	<0.11350	<0.09690
		REP5	0.08181	<0.11350	<0.09690
	1996	REP1	0.06146	<0.11350	<0.09690
		REP2	0.08055	<0.11350	<0.09690
		REP3	0.08516	<0.11350	<0.09690
		REP4	0.03227	<0.11350	<0.09690
		REP5	0.06266	<0.11350	<0.09690

We considered 1.63 for 1% and 1.36 for 5% significance level, respectively.

In the analysis of the results, we note that the null hypothesis that the sample comes from a normal population, is rejected in five cases for the significance level of 5% and only in one case to significance level of 1%, which is not altogether surprising since we are facing a total of 42 cases.

If we consider a binomial distribution with $n = 42$ and $x = 5$ then it's expected that 5% of observations fall outside the standard of reference, in our case the normal distribution.

6 Considerations and remarks

The vegetative reproduction seems to guarantee the homogeneity of genotypes. So in a given year and local the productions of the same genotype should be distributed normally. Not always this happens because when we studied the 42 repetitions of genotypes: Aragonês, Trincadeira Preta and Touriga Nacional four cases were found where the theoretical model did not fit significantly. In these four cases it was possible to fit the data a mixture of two normal distributions.

We consider important in future work using the techniques of ANOVA to estimate the variance components internal to the genotypes and between genotypes.

References

1. Bhattacharya, C.G. (1967) A simple method for resolution of a distribution into its Gaussian components. Biometrics 2, p.115-135.
2. Cassie, R.M. (1954) Some uses of Probability Paper for the Graphical Analysis of Polymodal Frequency Distributions. Austral J. Marine and Freshwater Res. 5, p.513-522.
3. Dempster, A.P, Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J.R. Statist. Soc. B 39, p.1-38.
4. Fowlkes, E.B. (1979) Some Methods for Studying the Mixture of two normal (lognormal) distributions. JASA 74, p.561-575.
5. Harding, J.P. (1948) The Use of Probability for the Graphical Analysis of Polymodal Frequency Distributions. J. Marine Biol. Assoc. U.K. 28, p.141-153.
6. Oliveira, A. (1999) Misturas de Normais: Uma Aplicação. Tese de Mestrado. FCT/UNL.
7. Pearson, K. (1894) Contributions to the Mathematical theory of evolution. Phil. Trans. A. 185, p.71-110.
8. Quandt R. E. & Ramsey, J. B. (1978) Estimating Mixtures of Normal Distributions and Switching Regressions (with discussion). JASA, 73, p.730-752.
9. Teicher, H. (1963) Identifiability of Finite Mixtures. Ann. Math. Statist., 34, p. 1265-1269.