https://doi.org/10.26219/heal.aueb.6203

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS
MSc IN STATISTICS

MODEL BASED CLUSTERING FOR YOUNG STAR CLUSTERS

By
Efthymia Derezea
Supervisor: Dimitris Karlis

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens,Greece
September 2018

https://doi.org/10.26219/heal.aueb.6203

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΑΝΑΛΥΣΗ ΣΕ ΟΜΑΔΕΣ ΜΕ ΤΗ ΧΡΗΣΗ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ΓΙΑ ΟΜΑΔΕΣ ΝΕΑΡΩΝ ΑΣΤΡΩΝ

Ευθυμία Δερεζέα
Επιβλέπων Καθηγητής: Δημήτρης Καρλής

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Σεπτέμβριος 2018

https://doi.org/10.26219/heal.aueb.6203

https://doi.org/10.26219/heal.aueb.6203

ACKNOWLEDGMENTS

I would like to thank my Supervisor Dimitris Karlis, for his support and comments regarding this thesis. Also thank you to all the people that supported me throughout this year, including my professors, my friends and family.

https://doi.org/10.26219/heal.aueb.6203

ABSTRACT

Model based clustering is a method of fitting finite mixtures models to data in order to identify clusters. It is widely used in many fields and has a variety of applications,one of those is the identification of young star clusters. We will examine four different models multivariate Gaussian, t, Skew Normal and Skew t mixtures and a method that suggests combining normal components to form a cluster. We will analyze their characteristics and we will then apply them to a real dataset of young stars in order to evaluate their performance in this specific problem.

https://doi.org/10.26219/heal.aueb.6203

Contents

https://doi.org/10.26219/heal.aueb.6203

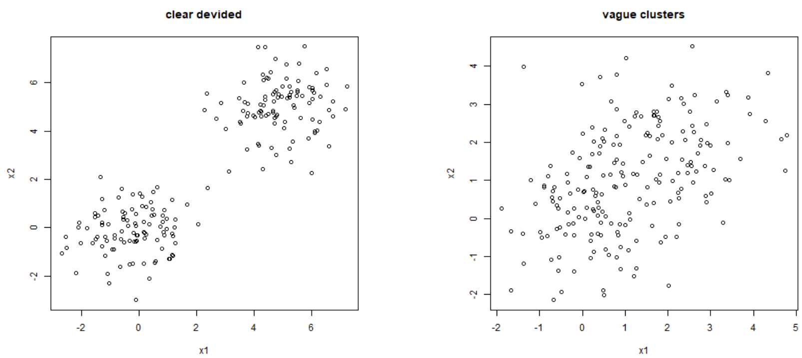
1 Introduction

Clustering is an important method of unsupervised learning in Statistics. Its main goal is to try and divide a number of observations into subgroups or "clusters". There is no unique way of separation, but the main idea is that information from available variables is used in order to categorize the observations. More precisely, we try to split the data into groups that have the following characteristics: First of all, the observations within a cluster or group should be as homogeneous as possible. Observations between clusters however, should look quite different. Figure 1 shows this more analytically.

Clustering has gained a lot of attention over the last years due to its various fields of applications. It is widely used in biology to cluster genes with similar characteristics for example. In marketing to categorize target groups. Many companies group their clients. It is a useful tool in Archeology. Another example is Astronomy, where clustering has many applications, for example it is used to determine galaxy or star clusters. The main application of this thesis is relevant to this. Almost every scientific field uses clustering for various purposes, so its importance is clear.

Defining clustering is not an easy task, since there is no single algorithm to work with. There exists a huge variety of approaches, some of them are completely algorithmic and others are based on statistical models. It is important to understand that there isn't a technique that surpasses all the others. Every approach has pros and cons, and it depends on the situation and the problem which is more suited to use every time. Some of the most famous approaches are the following: Hierarchical clustering, k-means algorithm, model based clustering and many more. Both Hierarchical clustering and k-means are algorithmic procedures and result into a hard clustering. Model based approach on the other hand, results into soft clustering giving the probabilities for an observation to belong into a cluster.

Unlike other methods, model based clustering allows for inferential tools. Having a model means, we have a likelihood, so we can use many of the established methods for example, goodness of fit, uncertainty, information criteria to decide amongst possible models etc. What is more, some of the other methods mentioned before, such as k-means are special cases of model based clustering. Having all these in mind is easy to understand the importance of probability models in clustering[12],[6]. Finite mixture models constitute the basic idea in order to apply the above, such models for clustering have been proposed many years now. When it comes to clustering we need to know how good our solution is, how many clusters or groups fit better the data. The answer to all these questions comes in a more natural way when model based clustering is used, it is also possible to deal with outliers.



(a) distant clusters

(b) overlaping clusters

Figure 1: In both cases we have data generated from a multivariate normal distribution. In the first case we have a group with mean (0,0) and (5,5) while in the second case the second group has mean (2,2). The variance matrix is the same for both. We notice that in some cases the groups are rather obvious while in other cases the distinction is not that easy.

The aim of this thesis is to gather and represent some of the basic mixture models and other techniques for model based clustering. Also, all these models will be used in an application with real data from the MYStIX project with young stars in nearby star formation regions (Kuhn et al. 2014) in order to evaluate their performances and make comparisons.

Specifically, the thesis is organized as follows. In Section 2, we will discuss what is a finite mixture model, how we determine them and their importance in clustering. Furthermore, we will refer to the basic idea of expectation-maximization (EM) algorithm, which is used in order to estimate the parameters of the models we use. In Section 3, we will examine more analytically the most famous or most used mixture models in clustering, but we will see in detail some of them. The basic theory will be explained and then there will be some simulation examples for illustration purposes.

First, the well known multivariate Normal mixture model will be represented, we will see analytically the model, the way we estimate the parameters, some simulated examples and we will refer to the packages available in R in order to implement these. Then, in sections 3.2 -3.4 we will discuss the basic variants of normal mixtures. Precisely, we will see in detail the multivariate t distribution and the multivariate Skew Normal and Skew t distributions. Other approaches in model based clustering are also discussed. In section 3.5 the methodology of combing mixtures components is represented. However, we will see only the merging of normal components.

In section 4, we will use all the above mentioned in real data. The application has to do with star clusters. First there will be a brief introduction to the data

https://doi.org/10.26219/heal.aueb.6203

and their purpose. We will apply the models to see their performance in this particular problem. We will compare the models and the approaches, using the gap statistic or the Wilk's Lamda and we will try to evaluate how similar the results we are getting are, using the ARI(Adjusted Rand Index). We are also going to see how well these models perform, in a small and a large number of groups case, since in this type of data we may deal with regions with a small number of groups for example 4-6, or with regions with up to 20 groups (subclusters of stars in particular) or more. Finally, in section 5 there will be a brief discussion for other approaches someone could use and a summary of the basic results and conclusions.

https://doi.org/10.26219/heal.aueb.6203

2 Mixture models

The basic idea in model based clustering, is that each different group corresponds to a mixture component. In order for this to become clear we need to understand what a finite mixture model is [26]. Generally, a finite mixture density is defined as follows:

$$p(x_i|\Theta_g) = \sum_{g=1}^{G} \pi_g p_g(x_i|\Theta_g)$$

By Θ_g , we denote a vector containing all the parameters for estimation. With π_g we denote the probability that an observation is "coming" from the g group with density $p_g(x_i|\Theta_g)$, or differently π_g are the proportional weights such that $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$. Figure 2 can provide more insight regarding a finite mixture density and π_gs' role. Seeing these, it is very natural for such models to be used in the clustering framework.

Having determined the model in this general form we understand, that estimating its parameters, is not a straightforward task since its log likelihood form is rather complicated. In order to estimate the parameters Θ_g and the weights π_g we should use the expectation-maximization algorithm (EM). The EM algorithm is a tool that allows us to estimate parameters in cases of missing data. It is widely used in many problems such as when dealing with grouped data, missing or truncated data etc. Generally, suppose that we have a set of data Y and we have a likelihood of the form $L(\theta|Y)$ that we wish to maximize it in order to obtain the estimators for the parameters Θ. If this likelihood is too difficult to do so (which is the case when dealing with mixture models), we add information with the form of missing data Z in order to obtain a likelihood easier to maximize, of the form $L(\theta|Y, Z)$.

Specifically, in its very general form EM is a two-step algorithm. In the first step (E-step or Estimation step), we have to estimate the missing data Z, from the observed likelihood $L(\theta|Y)$. First we define the following function given a current estimate $\theta^{(k)}$:

$$Q(\theta, \theta^{(k)}) = \int_Z \log L(\theta|Y, Z) P(Z|\theta^{(k)}, Y) dZ = E(\log L(\theta|Y, Z))$$

where $P(Z|\theta^{(k)}, Y)$ is the posterior distribution of Z. At E-step our goal is actually to calculate this Q($\cdot$) function.

Then proceeding to the second step (maximization or M step), we maximize the $L(\theta|Y, Z)$ likelihood with respect to θ, using the estimation for Z from the E-step, or in a formal way we maximize the function $Q(\theta, \theta^{(k)})$ with respect to θ. In many cases the ML estimators needed for M step are already known, but there are also instances, where the estimators are computed numerically,

https://doi.org/10.26219/heal.aueb.6203

using Newton Raphson for example. We start with some initial values and then repeat the above steps until a "termination" condition is satisfied. This condition could be based either on the likelihood or on the parameters and indicates lack of progress. A widely used stopping criterion for example, based on the likelihood is to terminate the algorithm when:

$$\frac{|L^{(k+1)} - L^{(k)}|}{|L^{(k+1)}|} < e^*$$

Where $L^{(k)}$ is the likelihood at the k-th iteration. We generally choose to stop when the difference between two iterations is very small(i.e $e^* = 10^{-6}$). It is proved that at each iteration the likelihood increases.

Moving on to the problem we are facing in clustering, we need to use EM in order to estimate the probabilities π_g as well as the other parameters. The log-likelihood of the data is given as follows:

$$l(\Theta, X) = \sum_{i=1}^{n} \log(\sum_{g=1}^{G} \pi_g p_g(x_i|\theta_g))$$

As we notice this likelihood is difficult to maximize as it is that's why we need to use EM. This problem would be easier if we knew in which group every observation belongs to. If that was the case, the parameters estimation would be easy and straightforward. Having these in mind, the idea is to augment our data with some unobserved variables Z_{ig}. We will suppose that these variables, will indicate in which cluster each observation belong to. These variables will take values zero or one. If for example, $Z_{ig} = 1$ this means that the i-th observations is from the g-th group, and it would be zero everywhere else. The new log-likelihood with the augmented data is given by:

$$l(\Theta|X, Z) = \sum_{i=1}^{n} \log(\sum_{g=1}^{G} (\pi_g p_g(x_i|\theta_g))^{z_{ig}}) =$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{G} (z_{ig} \log \pi_g) + z_{ig} \log(p_g(x_i|\theta_g))$$

By estimating the Z_{ig}s we know in which cluster each observation belongs to, so the maximization is easy as we can see from this complete data likelihood. We should also notice that p_g can be any distribution and it isn't obligatory for all components to be from the same. The procedure that we follow is exact the same described previously.

At this point it is important to mention that EM algorithm often fails to identify the true maximum and sometimes it gets trapped in local maxima. It is easy to understand if something like this is happening if we start our algorithm from different points and get different solutions. This makes clear the fact that choosing good initial values is a crucial part of the estimation. There are many strategies that one could follow for the initial values. First of all if prior

https://doi.org/10.26219/heal.aueb.6203

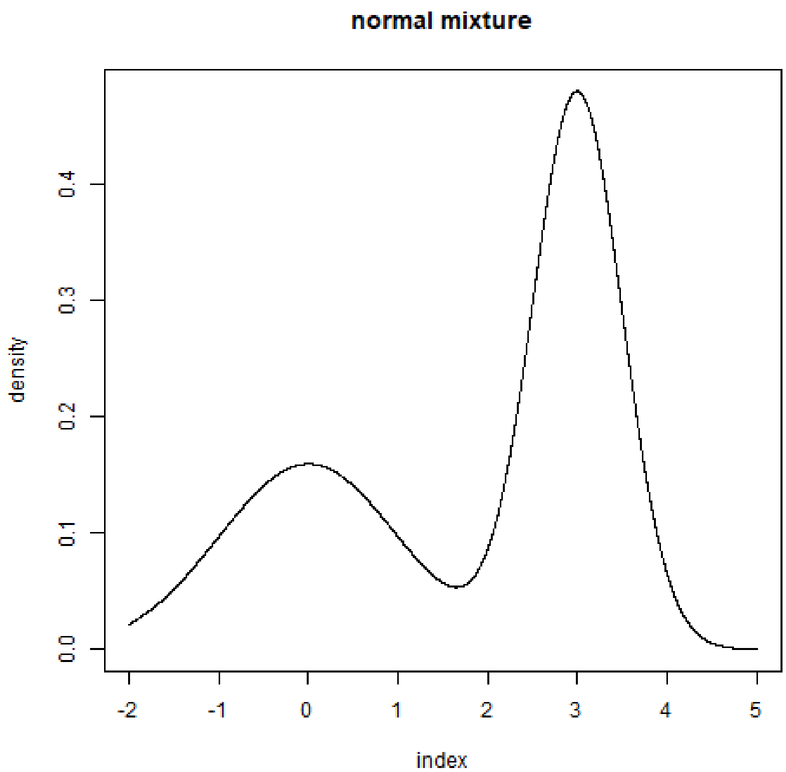


Figure 2: The density of a mixture of two normal distributions. Some observations are from the one subpopulation with probability π , while the others are from the other subpopulation with probability $1-\pi$. Sometimes the distinction between groups is not so obvious.

https://doi.org/10.26219/heal.aueb.6203

knowledge regarding the clusters of the data exists it would we useful to use it for initialization. Furthermore, one could use as initial values the centers of other clustering techniques such as the hierarchical clustering or the k-means algorithm. In general it is important to be sure that we have covered the hole space of our data, so another strategy would be to run EM many times, using different initial values each time and choose the better solution.More details for the initial values on EM can be found at [16], [8].

Having a mixture model it is also easy working this way to estimate the optimal number of clusters in a very natural way, comparing with other methods which is not easy. Information criteria such as BIC can be used to choose between possible numbers of clusters. Thus, this is the basic idea used in mixture models for model based clustering. Changing the distribution used each time can give us different results and choosing the appropriate can be a difficult task.

https://doi.org/10.26219/heal.aueb.6203

3 Some models and alternative methods.

In this section we will present basic models for multivariate clustering and alternative techniques. There is a huge variety of models to choose between. The most common is the finite mixture of multivariate Normal distributions. After this there were attempts to use distributions that would capture better non normal characteristics such as fat tails or asymmetry. Multivariate t distribution has been widely used [2] as well as many variants of skew normal and skew t, for example [20]. Except from these, other models have been proposed and used for clustering such as the normal inverse Gausian [15] or the shifted asymmetric Laplace [29]. Different approaches have been also proposed such as merging components of mixtures in order to form a group [7] ,[14] , as well as the use of copulas instead, in order to capture more peculiar shapes [17].

Possible Models	
Models/Approaches	**attribute for**
Mult. Normal mixtures	normality
Mult. t mixtures	fat tails
Mult. Skew Normal mixtures	asymmetry
Mult. Skew t mixtures	fat tails, asymmetry
Normal Inverse Gaussian mixtures	fat tails, asymmetry
Shifted asymmetric Laplace	asymmetry
Merging mixture components	peculiar shapes
Mult.copulas mixtures	peculiar shapes

Table 1:Possible models and techniques

 https://doi.org/10.26219/heal.aueb.6203

3.1 Multivariate Normal Mixtures

Multivariate normal distribution is most typically used in model based clustering. The idea discussed in the previous chapter is used here in exactly the same way, the only thing that changes, is that in the mixture model formula we replace $p_g(x|\theta_g)$, with the multivariate normal distribution. We assume, that every cluster from our data is coming from a multivariate normal distribution. The multivariate normal distribution density is given by:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(x-\mu)^{'}\Sigma^{-1}(x-\mu))$$

x is a vector p x 1, μ is the mean vector again with dimensions p x 1, and Σ is a variance-covariance p x p matrix.

To begin with, we need to determine the finite mixture of multivariate Gaussian distributions. The density is given, according to the previous mentioned, by the following formula:

$$f(x) = \sum_{g=1}^{G}\pi_g(2\pi)^{-\frac{p}{2}}|\Sigma_g|^{-\frac{1}{2}}exp(-\frac{1}{2}(x-\mu_g)^{'}\Sigma_g^{-1}(x-\mu_g))$$

where G denotes the number of components (or groups), μ_g is the g-th component mean and Σ_g is the g-th component covariance matrix.

In order to estimate the parameters we will use the EM algorithm as mentioned previously. We will again augment our data with the Z_{ig} variables. The likelihood would be of the form:

$$L = \prod_{i=1}^{n}\prod_{g=1}^{G}(\pi_g f(x_i|\mu_g,\Sigma_g))^{z_{ig}}$$

where f is the multivariate normal distribution.

The procedure, remains the same. First, we should appoint initial values to all the parameters we need to estimate(i.e. $\pi_1,...,\pi_g,\mu_1,...,\mu_g,\Sigma_1,...,\Sigma_g$). At E-step we need to estimate the Z_{ig}'s. In order to do so we calculate,

$$z_{ig} = E(Z_{ig}) = \frac{\pi_g f(x_i|\mu_g,\Sigma_g)}{\sum_{g=1}^{G}\pi_g f(x_i|\mu_g,\Sigma_g)} =$$

$$= \frac{\pi_g(2\pi)^{-\frac{p}{2}}|\Sigma_g|^{-\frac{1}{2}}exp(-\frac{1}{2}(x_i-\mu_g)^{'}\Sigma_g^{-1}(x_i-\mu_g))}{\sum_{g=1}^{G}\pi_g(2\pi)^{-\frac{p}{2}}|\Sigma_g|^{-\frac{1}{2}}exp(-\frac{1}{2}(x_i-\mu_g)^{'}\Sigma_g^{-1}(x_i-\mu_g))}$$

With the estimations of the z_{ig}s we can now proceed to the second step of the algorithm. The complete log-likelihood, as we saw before, is easy to maximize.

https://doi.org/10.26219/heal.aueb.6203

In the case of multivariate normal distribution the ML estimators can be found in close form. Using the following equations we update the parameters.

$$\pi_g^{(new)} = \frac{\sum_{i=1}^{n} z_{ig}}{n}$$

$$\mu_g^{(new)} = \frac{\sum_{i=1}^{n} z_{ig} x_i}{\sum_{i=1}^{n} z_{ig}}$$

$$\Sigma_g^{(new)} = \frac{\sum_{i=1}^{n} z_{ig}(x_i - \mu_g^{(new)})'(x_i - \mu_g^{(new)})}{\sum_{i=1}^{n} z_{ig}}$$

We repeat steps E and M, until the chosen stopping criterion is met.

At this point, it is important to mention that we can choose between a range of models of different shape, volume or orientation so the characteristics of the data are fitted better. This can be done if we rewrite the component covariance matrix as follows:

$$\Sigma_g = \lambda_g D_g A_g D_g'$$

With this eigen-decomposition of the Σ_g matrix we can choose between different parsimonious models [27], constraining each time some of the parameters. The parameter λ_g explains the volume, matrix D_g contains the eigenvectors and explains the orientation, finally A_g is a diagonal matrix with the eigenvalues that determines the shape. So, depending on the choice of the above we can search for clusters that for example all have the same shape, volume and orientation to situations where every cluster has different values for the above.

Having all these in mind it is easy to choose the best model and the most suited number of clusters using the Bayesian Information criterion. The case with the smallest number for BIC would be considered the most appropriate. Other criteria can also be implemented, for example the normalized entropy criterion or the integrated completed likelihood (more details could be found at [11])

All these can be easily implemented in R using ready packages such as the "mclust" or the "pgmm". In order to understand clearer the above mentioned some illustrations with simulated data will help. We will examine 6 cases in order to asses its performance and obtain a better understanding. For the clustering the "mclust" package in R will be used.

Specifically, the cases, will include data simulated from the multivariate normal, the multivariate t, multivariate skew normal and skew t, as well as a case were groups from all distributions will be used. The purpose for this is to obtain a clearer image of how model based clustering works and also an insight in how well multivariate normal will fit different types of data.

In Figure 4 we can see the groups that were identified using the mclust package. As it was expected, using multivariate normal model found perfectly the clusters in the first case where normal simulated data were used. However, when the data were generated from other distributions such as multivariate t or multivariate skew normal distributions more groups were needed in order to fit the data properly. In the fifth case were groups from various distributions

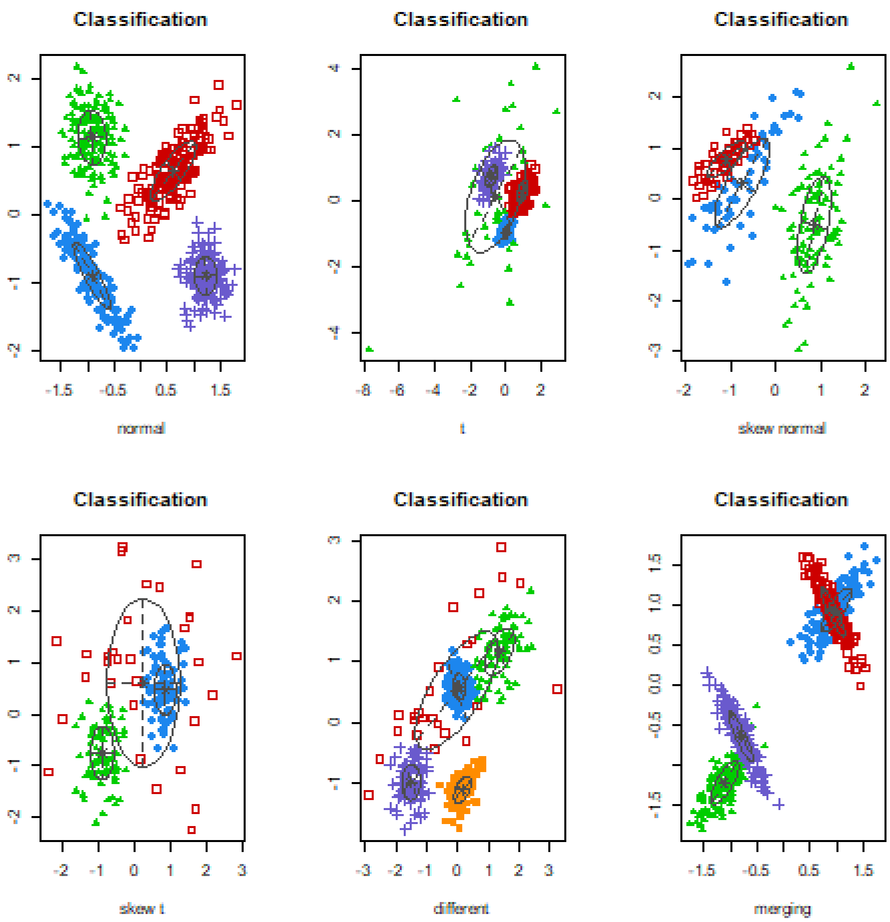


Figure 3: Clustering using Gaussian mixtures for 6 cases of simulated data.

were used again an extra cluster was fitted but as a result it wasn't that bad. Next we use adjusted rand index, which is a measure that shows us how similar these results are, compared to the true labels of the data. The ARI is defined as follows,

$$ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

and its value for all cases are the following: 0.99 for the normal case, 0.75,0.72, 0.71 for t and the skewed distributions and 0.91 for the fifth case with all the distributions.

https://doi.org/10.26219/heal.aueb.6203

3.2 Multivariare t Mixtures

Sometimes, the data we are using might deviate from the normality assumption. In such cases, other models have been proposed in an attempt to fit better the data. A very famous alternative is using mixtures of the multivariate t distribution. Using this distribution could help capture non-normality characteristics of the data such as "fat-tails".

In order to understand better the properties of this model and the way it works, we should see how it behaves in the univariate case. First of all t distribution's density is given by:

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}, x \in \mathbb{R}, \nu > 0$$

where ν denotes the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The shape of this distribution is similar to that of the normal, but it is more flat with wider tails. As parameter ν increases, t distribution gets closer to Normal. Figure 4 illustrates analytically this behavior. Another important characteristic we should mention is that we can represent t by the following way,

$$T = \frac{Z}{\sqrt{V/\nu}}$$

where Z follows the Standard Normal distribution, $Z \sim N(0,1)$ and $V \sim X_\nu^2 \equiv \Gamma(\nu/2, 2)$.

Moving on,the multivariate t density is given by:

$$f(x|\mu, \Sigma, \nu) = \frac{\Gamma(\frac{\nu+p}{2})|\Sigma|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})(1+\frac{(x-\mu)^\tau\Sigma^{-1}(x-\mu)}{\nu})^{\frac{\nu+p}{2}}}$$

where μ is a px1 vector with the means, Σ is a p x p scale matrix and with ν we denote the degrees of freedom. $\Gamma(\cdot)$ is the Gamma function.Figure 5 provides more details regarding the behavior of the multivariate t distribution. The density, of the finite mixtures of multivariate t distributions, is defined by the same way again as in the normal case. We only substitute the above density function. The model would be defined as follows:

$$f(x|\mu_g, \Sigma_g, \nu_g) = \sum_{g=1}^{G} \pi_g \frac{\Gamma(\frac{\nu_g+p}{2})|\Sigma_g|^{-\frac{1}{2}}}{(\pi\nu_g)^{\frac{p}{2}}\Gamma(\frac{\nu_g}{2})(1+\frac{(x-\mu_g)^\tau\Sigma_g^{-1}(x-\mu_g)}{\nu_g})^{\frac{\nu_g+p}{2}}}$$

again μ_g is the mean vector for the g-th component, Σ_g is the scale matrix of the g-th component and finally ν_g is the degrees of freedom for the g-th group.

https://doi.org/10.26219/heal.aueb.6203

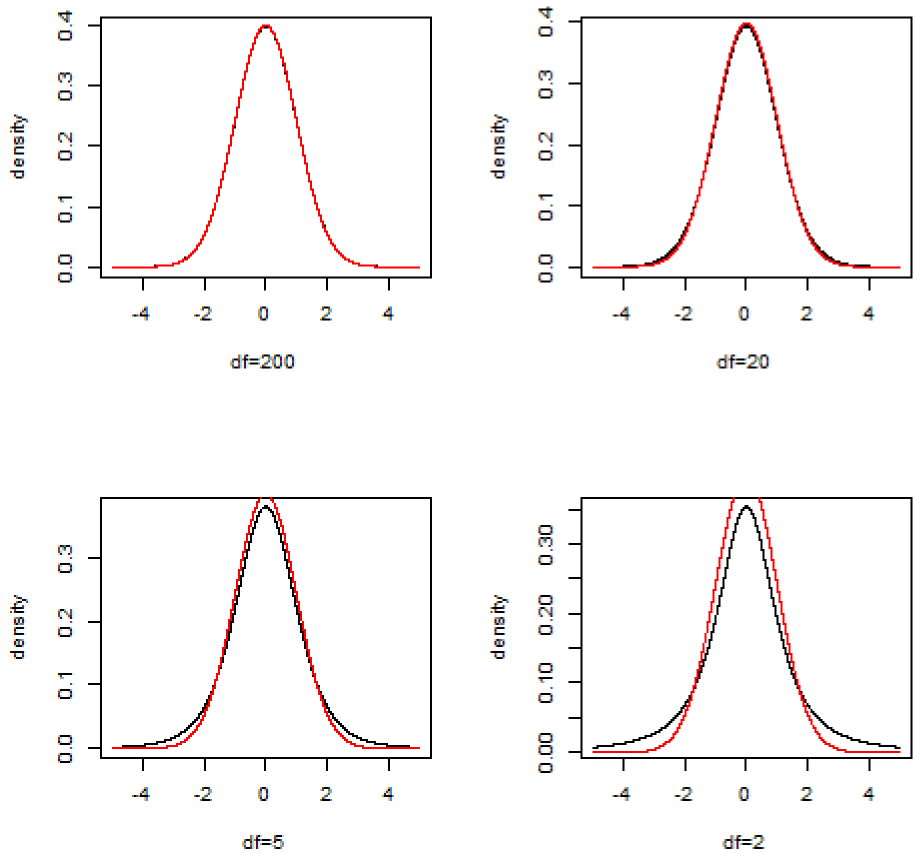


Figure 4: The t distribution for different degrees of freedom. With red line we see the density of the normal distribution. We notice that as the number of degrees of freedom increases t goes closer to normality.

https://doi.org/10.26219/heal.aueb.6203

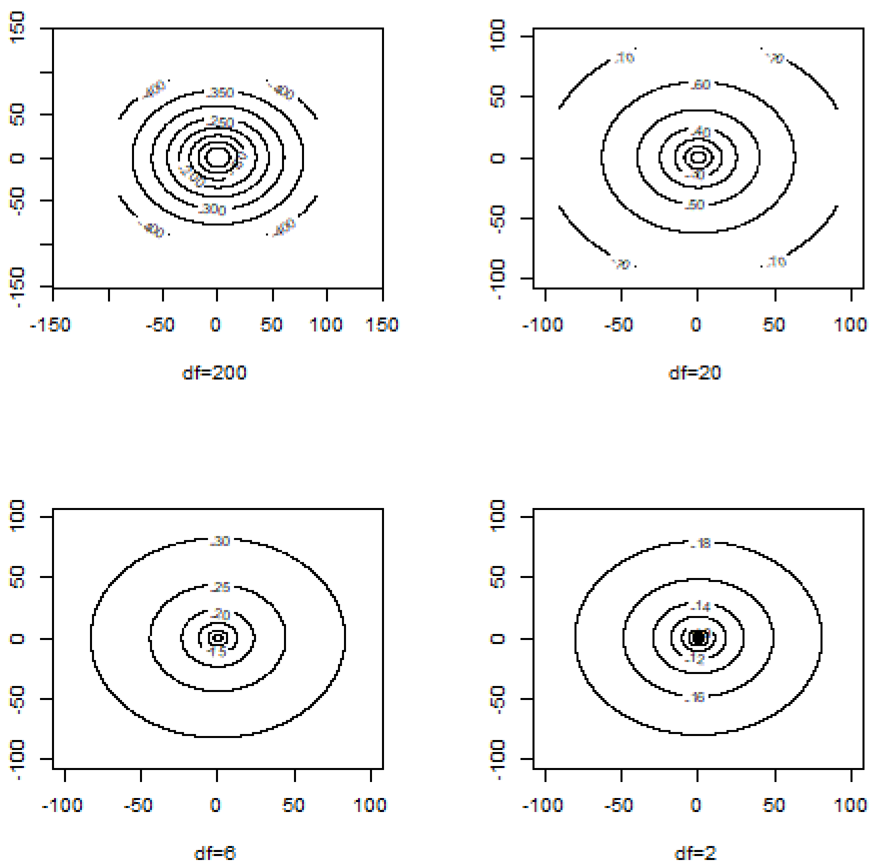


Figure 5: Contour plots for the bivariate t distribution for different degrees of freedom.

 https://doi.org/10.26219/heal.aueb.6203

Similarly with the Gaussian mixtures case, we can rewrite the scale matrix Σ_g as follows:

$$\Sigma_g = \lambda_g D_g A_g D_g'$$

The notation is the same. D_g is a matrix with the eigenvectors, A_g is a diagonal matrix with the eigenvalues and λ_g is a scalar controlling the volume. This decomposition of the scale matrix allows us to choose between a range of models restraining some of the parameters each time. In this case of multivariate t distribution taking into account and the degrees of freedom ν_g, we result into a 28 range of different models to choose between (of the same family).

In order to estimate the model parameters we have to use the EM algorithm[13]. At this point we should mention that the procedure we described previously is the most general form of the EM algorithm. There exist many variants. In this mixture problem, is used a variant of EM, called the expectation-conditional maximization (ECM) algorithm [28]. ECM is based on an algorithm called GEM. In this algorithm in M-step instead of maximizing the complete data likelihood we constantly increase it (this may lead to convergence issues.) ECM replaces the M step with a series of conditional maximization steps, called CM-steps. Each of these steps maximizes the complete log-likelihood with respect to the parameters but at each time some of them are fixed to their previous values, so each CM step increases the likelihood.

The procedure is similar. We have again to define the unknown variables Z_{ig} so that if an observation belongs to the g-th cluster $Z_{ig} = 1$ and otherwise zero. The complete data log-likelihood is given by:

$$l = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} log[\pi_g \gamma(u_{ig}|\nu/2, \nu/2) \phi(x_i|\mu_g, \Sigma_g/u_{ig})]$$

where $\gamma(\cdot)$ is the density function of Gamma distribution, and $\phi(\cdot)$ is the normal density. We notice that here, the analogous multivariate representation for t, mentioned in the univariate case is used (more details, [1]). As we can see we can use this hierarchical representation with Normal and Gamma distribution that allows us to use the EM as follows.

At the E step we have to estimate the variables Z_{ig} as well as the characteristic weights U_{ig} these could be thought as a metric of the influence each observation has in the estimation of the mean and the scale matrix. The expectations we need to calculate are the following.

$$E[Z_{ig}|x_i] = \frac{\pi_g f(x|\mu_g, \Sigma_g, \nu_g)}{\sum_{g=1}^{G} \pi_g f_g(x|\mu_g, \Sigma_g, \nu_g)}$$

$$E[U_{ig}|x_i, z_{ig} = 1] = \frac{\nu_g + p}{\nu_g + d(x_i, \mu_g|\Sigma_g)}$$

where $d(x_i, \mu_g|\Sigma_g)$ is the squared Mahalanobis distance between x_i and μ_g.

Having obtained the estimates of z_{ig}'s and u_{ig}'s which are the expected values calculated at the E step respectively, we move on to the CM steps. Here

https://doi.org/10.26219/heal.aueb.6203

we will conditionally maximize the complete log-likelihood with respect to the parameters. On the first CM-step we update the parameters π_g and μ_g using the following formulas.

$$\pi_g = \frac{\sum_{i=1}^{n} z_{ig}}{n}$$

$$\mu_g = \frac{\sum_{i=1}^{n} z_{ig} u_{ig} x_i}{\sum_{i=1}^{n} z_{ig} u_{ig}}$$

On the second CM-step the scale matrix is updated, or better the decomposed elements λ_g, D_g and A_g (more details at[3]). As far as the degrees of freedom concerned their updated by either numerical methods or by some closed form approximations [2].

These steps are repeated until the chosen stopping criterion is satisfied. Again, we can use the likelihood and the relevant information criteria in order to choose between models and number of clusters. In R these are implemented with the package "teigen".

In order to acquire a better understanding of the use, of multivariate t mixtures in clustering we will use again some simulation examples. The simulated datasets will be the same with those used in the section with the normal mixtures in order to be able to see the differences between the two models.Figure 6 shows the six cases likewise with the previous example with the use of "teigen" package in R.

To begin with, we notice that with the use of multivariate t mixtures instead of normal we were able to identify the correct number of cluster in almost all cases except for the last one. This perhaps due to the fact that since with the t distribution we can capture the fat-tails characteristics there is no need for an extra cluster (which was the case with normal). What is more, this model seems to fit better the data than the normal in all the cases. The Adjusted Rand Indexes seem to be also better with exception for the first case (normal data) and the last where the values were the same. Analytically, the ARIs are 0.85 for the t, 0.96 for the skew normal, 0.88 for the skew-t data and 0.88 for the fifth case with the different distributions. We notice that the results are quite improved for all the cases except the fifth, were the equivalent ARI for the normal model was 0.91.

https://doi.org/10.26219/heal.aueb.6203

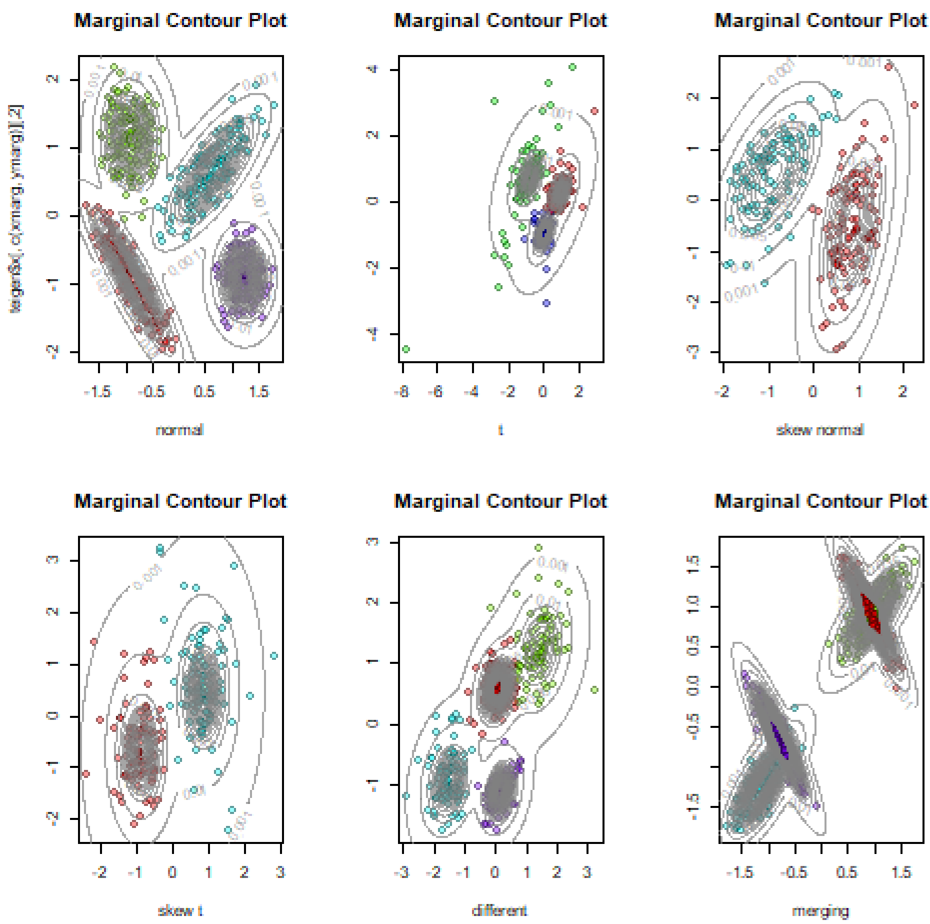


Figure 6: Clustering using multivariate t mixtures for 6 cases of simulated data.

3.3 Multivariate Skew Normal Mixtures

In many cases, data can have various features that need to be captured, from a suitable model in order to be fitted correctly. As we saw in sections 3.1 and 3.2 the Gaussian Mixtures models are restricted to provide us with elliptical-shaped clusters. What is more, we saw that when using normal mixtures we tend to overfit the data in order to capture the characteristics, so we end up with solutions that indicate a larger number of clusters than the actual one. On the other hand, we saw that multivariate t mixtures perform better in such situations being able to capture the heterogeneity in the data. However, there are still situations that the above models can not handle properly, for example when it comes to asymmetric features. Mixtures of multivariate skew normal distributions can provide an alternative when it comes to fitting such data.

To begin with, the skew normal distribution is a class of densities, containing the Normal distribution, depending on another shape parameter. Specifically, if X is a variable following the skew normal distribution, its density function is derived as follows:

$$f(x; \lambda, \mu, \sigma^2) = 2\phi(x)\Phi(\lambda x)$$

where X is real valued, $\phi(\cdot)$ is the standard normal density, and $\Phi(\cdot)$ is the standard normal distribution function. The shape parameter λ takes values in $(-\infty, +\infty)$, if $\lambda = 0$ we have the normal distribution. If $\lambda > 0$ then the distribution is positive (left) skewed and when $\lambda < 0$ in negative (right) skewed. the larger the absolute value of λ, the larger the skewness of the distribution. Figure 7 provides more insight regarding the shape of the distribution.

Before we move on to the multivariate case, we should mention an important proposition which states that if (X_1, X_0) is a bivariate Normal r.v. with standar dized marginals and correlation $\delta(\lambda)$, then the conditional distribution of X_1 given $X_0 > 0$ is a Skew Normal distribution with skewness parameter λ, where δ is defined as follows:

$$\delta = \frac{\lambda}{(1 + \lambda^2)^{1/2}}$$

These are the fundamentals for the multivariate case as well[23]. The multivariate Skew normal distribution has been studied from many perspectives, many approaches have been used, and there exists a quite large number of variants. Here we will use and discuss about the fundamental skew normal distribution or FUSN using the notation from [20]. It is a more generalized form of the classical multivariate Skew Normal. From this family we can derive four types of distributions corresponding to the following forms: restricted, unrestricted, extended and generalized [4]. This distribution is generated in accordance to the previous by conditioning a multivariate normal on another variable, that could be either univariate or multivariate. Specifically,let $X_1 = (X_1, ..., X_p) \sim$

https://doi.org/10.26219/heal.aueb.6203

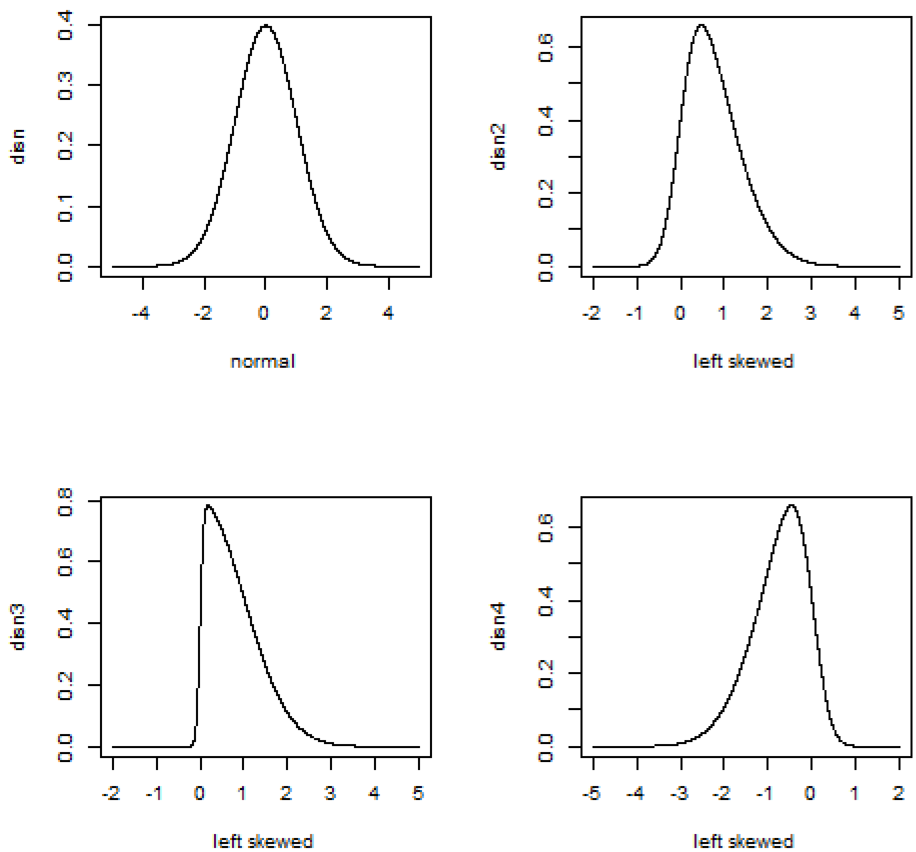


Figure 7: The Skew Normal Distribution.The top left graph shows the density function of a skew normal distribution for $\lambda = 0$, which is actually the standard normal distribution. On the top right we see the skew normal distribution for $\lambda = 3$, looking at the bottom left graph which is a skew normal for $\lambda = 15$, we see that as λ gets larger the skewness is more intense. Finally, in the bottom right graph we have a skew normal distribution for $\lambda = -3$.

https://doi.org/10.26219/heal.aueb.6203

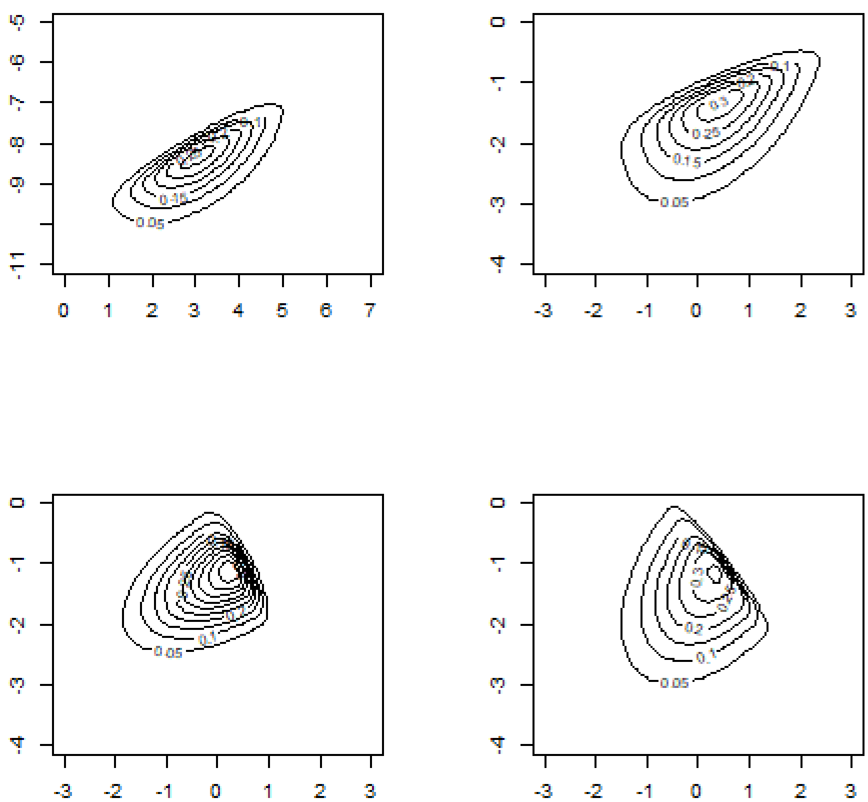


Figure 8: Contour plots of the bivariate Skew Normal distribution. Examples of possible shapes it can take changing the parameters.

https://doi.org/10.26219/heal.aueb.6203

$N_p(0,\Sigma)$ and X_0 a q-dimensional vector ,then $X = \mu + (X_1|X_0 + \tau > 0)$ has the fundamental skew normal distribution. Restricting the parameters of the distribution of $[X_1, X_0]'$ we can take one of the four forms mentioned previously,

$$\begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \tau \\ \mu \end{bmatrix}, \begin{bmatrix} \Gamma & \Delta^T \\ \Delta & \Sigma \end{bmatrix}\right)$$

where τ is a q-dimensional location parameter for the X_0, μ a p-dimensional vector, Γ a q x q scale matrix, Δ a p x q unrestricted matrix, and finally Σ is a q x q scale matrix as well.

If we choose X_0 to be univariate and follow a Normal distribution (so q=1), and $\tau = 0$ we are in the restricted case, if we keep the same restrictions but let X_0 to be multivariate (q=p),we have the unrestricted case. If the only restriction we pose is that X_0 is normally distributed, then we are in the extended case (here τ can be different from zero). Finally, if we let X_0 follow any distribution (so not only Normal), we have the generalized case. As mentioned before these are very general cases of the multivariate Skew Normal distribution, one can find many specific variants for example (Branco and Dey 2011,Azzalini 1985 ,Pyne et al.2009 ,Sahu et al. 2003). Figure 8 provides details regarding the shape of a multivariate skew normal distribution.

We should now continue to the finite mixtures of multivariate Skew Normal distributions. Using the previous notation the density of a mixture of g components is given as follows:

$$f(x;\Theta) = \sum_{g=1}^{G} \pi_g f_g(x;\Theta_g)$$

where, π_g is the probability of an observation coming from the g-th component, Θ is a vector containing all the unknown parameters, and $f_g(\cdot)$ is the multivariate skew normal density described above.

For the estimation of the parameters we use again the EM algorithm. The procedure is similar with the previous ones. We should augment our data with the random variable Z_{ig} ,where Z_{ig}=1 if the i-th observation belongs to the g-th component and takes the value 0 otherwise, estimate them at E-step and then by maximizing the complete-log likelihood at M-Step, obtain the estimates for the parameters. We repeat these steps until a termination criterion is satisfied as mentioned before. Analytically, as an illustrative example we can use the independent Skew normal distribution [10] which belongs to the restricted case.

First of all, we should note that the distribution of $X_i|Z_{ig} = 1$ is an independent Skew normal distribution, or $SNI_p(\mu_g, \Sigma_g, \lambda_g, \nu)$ and from the way Z_{ig} are defined, each $Z_i \sim Multinomial(1; \pi_1, ..., \pi_g)$. Moving on, the following hierarchical representation holds:

$$X_i|U_i = u_i, T_i = t_i, Z_{ig} = 1 \sim N_p(\mu_g + \Delta_g t_i, u_i^{-1}\Gamma_g)$$

$$T_i|U_i = u_i, Z_{ig} = 1 \sim HN(0, u_i^{-1})$$

$$U_i|Z_{ig} = 1 \sim H(u_i|v)$$

https://doi.org/10.26219/heal.aueb.6203

where, U is a positive random variable with distribution H(—ν), and ν is a scalar indexing H. $\Delta_g = \Sigma_g^{1/2}\delta$, $\Gamma_g = \Sigma_g - \Delta_g\Delta_g^T$, and as HN($\cdot$), is denoted the positive half normal distribution. The complete likelihood of the data is given by:

$$l = K + \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig}[log(\pi_g) - \frac{1}{2}log|\Gamma_g| - \frac{u_i}{2}(x_i - \mu_g - \Delta_g t_i)^T\Gamma_g^{-1}(x_i - \mu_g - \Delta_g t_i) + log(h(u_i|\nu))]$$

With K we denote a constant part independent of the parameters and thus not of direct interest. At the E-step of the algorithm the following conditional expectations are needed to be calculated:

$$\rho_{ig} = E[Z_{ig}|x_i]$$

$$\beta_{ig} = E[Z_{ig}U_i|x_i]$$

$$\xi_{ig} = E[Z_{ig}U_iT_i|x_i]$$

$$w_{ig} = E[Z_{ig}U_iT_i^2|x_i]$$

Here a variant of the EM algorithm is used, the ECME algorithm [25], which is a generalization of ECM algorithm that replaces CM steps that maximize the constrained actual likelihood. Specifically, at the first M-step at the k+1 iteration we update the parameters π_g, μ_g, Γ_g and Δ_g using the following formulas:

$$\pi_g^{(k+1)} = \frac{\sum_{i=1}^{n}\rho_{ig}^{(k)}}{n}$$

$$\mu_g^{(k+1)} = \frac{\sum_{i=1}^{n}(\beta_{ig}^{(k)}x_i - \xi_{ig}^{(k)}\Delta_g^{(k)})}{\sum_{i=1}^{n}\beta_{ig}^{(k)}}$$

$$\Delta_g^{k+1} = \frac{\sum_{i=1}^{n}\xi_{ig}^{(k)}(x_i - \mu_g^{(k+1)})}{\sum_{i=1}^{n}w_{ig}^{(k)}}$$

$$\Gamma_g^{(k+1)} = (\sum_{i=1}^{n}\rho_{ig}^{(k)})^{-1}\sum_{i=1}^{n}(\beta_{ig}^{(k)}(x_i - \mu_g^{(k+1)})(x_i - \mu_g^{(k+1)})^T - [(x_i - \mu_g^{(k+1)})(\Delta_g^{(k+1)})^T +$$

$$+\Delta_g^{(k+1)}(x_i - \mu_g^{(k+1)})^T]\xi_{ig}^{(k)} + \Delta_g^{(k+1)}(\Delta_g^{(k+1)})^T w_{ig}^{(k)})$$

At the second M-step at the k+1 iteration we update the ν by maximizing the actual log-likelihood with respect to ν:

$$\sum_{i=1}^{n} log(\sum_{g=1}^{G}\pi_g f_{g,SNI}(x_i|\mu_g^{(k+1)}, \Sigma_g^{(k+1)}, \lambda_g^{(k+1)}, \nu))$$

This is an example for a specific case of multivariate skew normal distribution. Depending on which variant we choose each time some of the steps may be different, but the main idea is the same. The implementation of EM algorithm

https://doi.org/10.26219/heal.aueb.6203

for mixtures of multivariate skew normal distributions is available in many packages in R such as the "sn" package, the "mixsmsn" ,"EMMIX-skew" and as special cases in the package "EMMIXcskew" [21] for example.

In order to evaluate this model and understand better its function we will use the same simulated data that were used in the previous sections. The data are simulated in order to represent six different cases, we will use the package "EMMIX-skew" in order to fit the skew normal mixtures to find the clusters.This package implements the restricted case of multivariate skew normal distribution and we have five different choices for the covariance matrix. Figure 8, illustrates the clustering obtained for the six cases with the use of this model.

We notice that in the case of the normal data, this model fits the data for 3 clusters (instead of 4 which is the actual number of clusters). In the t data, we have the poorest fit again the number of clusters found is one less from the actual number. we see that we have almost perfect results when we deal with skew normal data (as expected). The results seem also good in the fifth case with the data coming form different distributions, since it gives us clusters very close to the real ones. Again we are not able to capture groups such as in the sixth case. The Adjusted Rand Indexes for all the six cases are 0.52 for the normal case, 0.30 for the second case (data from the multivariate t distribution), 0.92 and 0.68 for the skew normal and skew t respectively, 0.95 for the case with the different distributions and finally 0.50 for the sixth case. We notice that the results compared with those from the t mixtures in the previous section are worse expect for the fifth case where the data are generated from different distributions. Compared to the results from the Gaussian mixtures, the results are better except for the first two cases.

https://doi.org/10.26219/heal.aueb.6203

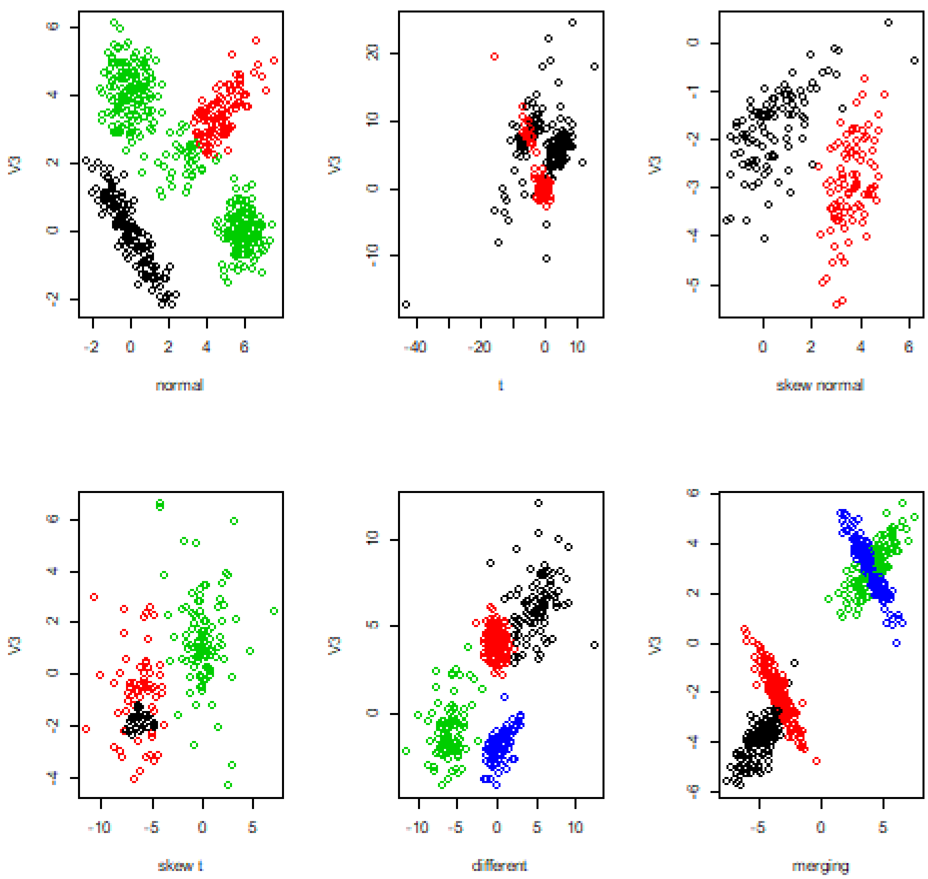


Figure 9: Fitting finite mixtures of multivariate skew normal distributions in six cases of simulated data.

https://doi.org/10.26219/heal.aueb.6203

3.4 Multivariare skew t mixtures

There are cases were the data have complicated features departing from normality, in which the above mentioned models cannot provide a satisfactory fit. Using finite mixtures of multivariate skew t distributions can provide us with flexible clusters, able to capture the "fat-tail" characteristic as well as the possible asymmetric nature of the data. In the previous sections we saw that using the multivariate t instead of the multivariate normal distribution helped us fit more appropriately data that had heterogeneous characteristic without overestimating the number of the groups. On the other hand using the multivariate skew normal distribution helped us to capture asymmetric characteristics however we still were unable to deal with the "fat-tails" problem. Having all that in mind, is clear that using a distribution able to capture both these characteristics could be extremely useful and perhaps could solve many of these problems.

To begin with, it is essential to first understand the basic characteristics of the univariate skew t distribution. Let X be a random variable represented as follows:

$$X = \mu + \sigma \frac{Z}{\sqrt{\tau}}$$

where, Z follows the skew normal distribution with skewness parameter λ, τ follows the Gamma distribution $\Gamma(\nu/2, \nu/2)$, and Z and τ are independent. Then the random variable X follows the skew t distribution, or X follows $ST(\mu, \sigma^2, \lambda, \nu)$, where μ is the location parameter of the distribution taking values in $(-\infty, +\infty)$, σ^2 is the scale parameter taking values in $(0, +\infty)$, λ is the skewness parameter as mentioned before taking values in $(-\infty, +\infty)$, and finally ν is the parameter denoting the degrees of freedom and it takes positive values. Similarly to the skew normal distribution as λ rises in absolute values the skewness of the distribution is greater ,positive values indicate left skewed distributions and negative right skewed. The degrees of freedom are similar to the case of t distribution as ν rises the distribution resembles more to the normal. Figure 10 provides more information regarding the skew t distribution and its behavior depending on the values of the parameters.

At this point it is important to mention that, the skew t distribution can be represented by the following hierarchical scheme:

$$X|\gamma, \tau \sim N(\mu + \delta\gamma, \frac{1-\delta^2}{\tau}\sigma)$$

$$\gamma|\tau \sim HN(0, \frac{\sigma^2}{\tau}; (0, +\infty))$$

$$\tau \sim \Gamma(\nu/2, \nu/2)$$

https://doi.org/10.26219/heal.aueb.6203

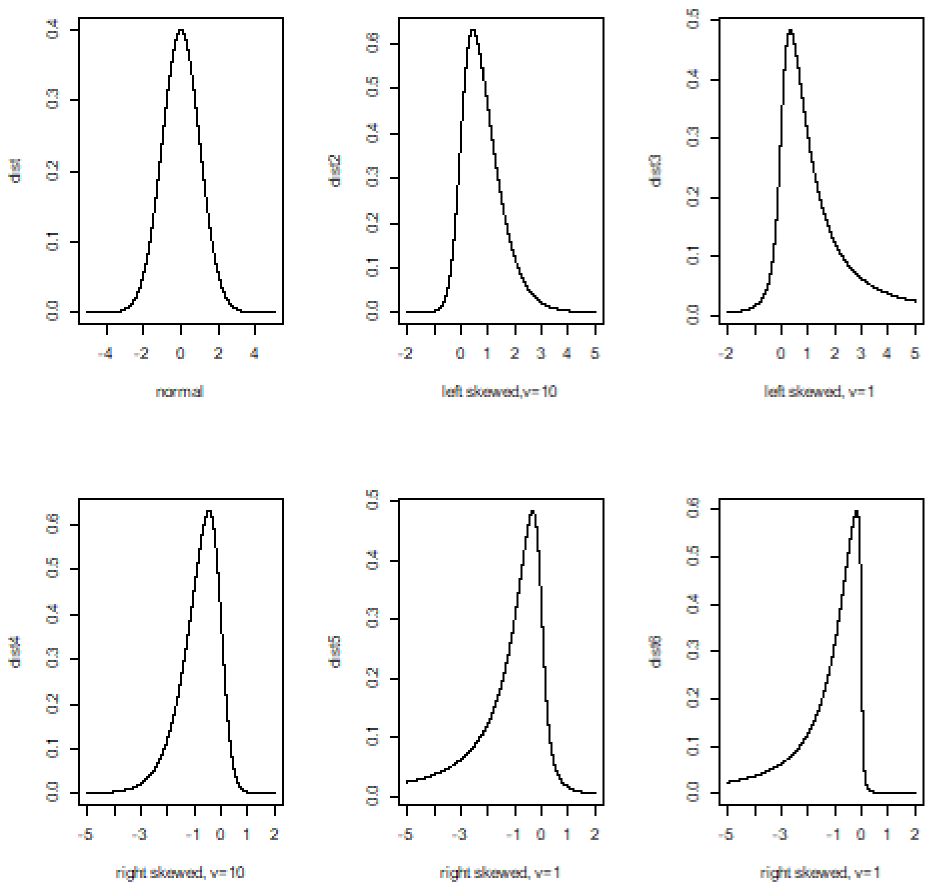


Figure 10: The Skew t distribution. The top left graph shows the density of a skew t distribution for $\lambda = 0$ and and degrees of freedom close to infinity, which is actually a normal distribution. The next two graphs illustrate the behavior of the skew t distribution as the number of degrees of freedom falls and the skewness parameter is equal to 3. In the first two graphs of the second row we notice the same thing the difference is that $\lambda = -3$, finally, in the last graph we see how the skewness changes as λ gets bigger ($\lambda = -15$).

https://doi.org/10.26219/heal.aueb.6203

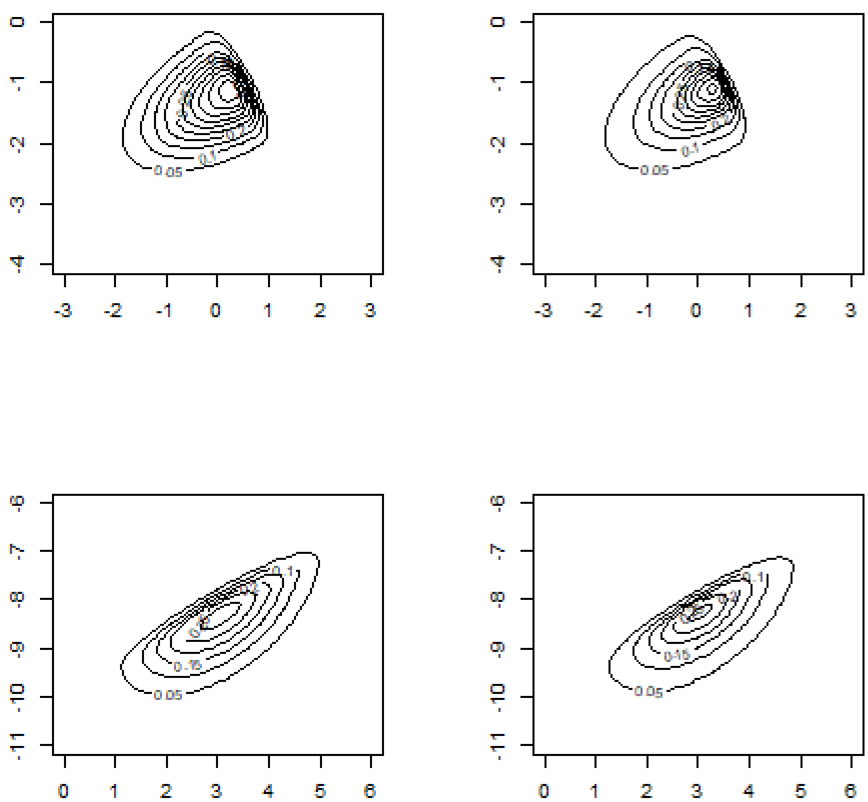


Figure 11: Contour plots for the bivariate skew t distribution. possible shapes for different values of the parameters.

https://doi.org/10.26219/heal.aueb.6203

This is useful because it helps us understand better the calculations needed in the multivariate extention that follows. More details can be found at [24].

Moving on to the multivariate t distribution, we should mention that again similar to the multivariate skew normal distribution, there exist many variants. In general these distributions can be categorized in four wide groups, namely as previously, as "restricted", "unrestricted", "extended" and "generalized". Most of the multivariate skew t variants are special cases of the fundamental skew t. Figure 11 provides more details regarding the behavior of multivariate skew t distribution.

Due to simplicity reasons, we will describe analytically the restricted case. As we said the number of variants even in one of the categories is huge there is for example the skew t distribution of Branco and Dey [9], Azzalini and Capitanio [5], the independent skew t ect. We will describe the general case given by Lee and MacLachlan. Following the same notation as in section 3.3 , if $X_1 = (X_1, ..., X_p)'$ and X_0 a univariate variable and the joint distribution of X_1 and X_0 is described as follows:

$$\begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \sim t_{p+1}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \delta^T \\ \delta^T & \Sigma \end{bmatrix}, \nu\right)$$

then if we condition on X_0 being positive we will obtain the restricted multivariate skew t distribution,specifically:

$$X = \mu + (X_1|X_0 > 0),$$

X has the multivariate skew t distribution. Here δ is defined the same way as a function of the scale parameter λ, Σ is a p x p scale matrix and with ν we denote the degrees of freedom. The variants mentioned above are obtained through the same mechanism, changing the parametrization. The second most simple family of multivariate skew t distributions is the unrestricted one. The only thing that changes here is that X_0 now is a p-dimensional random vector following a t distribution. At this point it is important to remember the hierarchical representation of the univariate case, because it can be extended to the multivariate case as well. We can write:

$$X = \delta|U| + U_0,$$

then

$$U_0|u, w \sim N_p(\mu + \delta|u|, \Sigma/w)$$
$$U|w \sim N(0, 1/w)$$
$$W \sim \Gamma(\nu/2, \nu/2)$$

This representation is important in order use the EM algorithm for parameter estimation.

Continuing, with the finite mixture model the basic formula is again the same.

$$f(x|\Theta) = \sum_{g=1}^{G} \pi_g f_g(x|\Theta_g)$$

https://doi.org/10.26219/heal.aueb.6203

where $f_g(\cdot)$ now denotes the multivariate skew t distribution. More precisely for the case of the restricted multivariate skew t distribution we can write:

$$f(x|\Theta) = \sum_{g=1}^{G} \pi_g MST(x|\mu_g, \Sigma_g, \delta_g, \nu_g)$$

In order to estimate the parameters of the model, we will use the EM algorithm [32]. Following the same basic idea we will again augment our date with some hidden or unobserved variables in order to reach a more manageable log-likelihood. First we will consider the variables U and W as latent in order to use the hierachical representation mentioned above. We will augment our data with the missing variables Z_{ig}, similarly to the previous sections Z_{ig}=1 if the i-th observation belongs to the g-th group, and 0 otherwise. Having all that in mind the complete log-likelihood is [19]:

$$l(|x,u,w,z) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} log(\pi_g) +$$

$$+\sum_{i=1}^{n}\sum_{g=1}^{G} \frac{z_{ig}}{2}[log|\Sigma_g|^{-1} - log(2\pi) + w_i(x_i-\mu_g-\delta_g u_i)^T \Sigma_g^{-1}(x_i-\mu_g-\delta_g u_i)]+$$

$$+\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}[-\frac{1}{2}[(p-1)log(w_i)+w_g u_i^2]+\frac{\nu_g}{2}[w_i - log(\frac{\nu_g}{2})]+log\Gamma(\frac{\nu_g}{2})+(\frac{\nu_g}{2}-1)logw_i]$$

where p is the dimension of X.

At the E-step of the algorithm we will need to calculate the following expectations.

$$\tau_{ig} = E[Z_{ig}|x_i]$$

$$e_{1,ig} = E[log(W_i)|x_i, Z_{ig}=1]$$
$$e_{2,ig} = E[W_i|x_i, Z_{ig}=1]$$
$$e_{3,ig} = E[W_i U_i|x_i, Z_{ig}=1]$$
$$e_{4,ig} = E[W_i U_i^2|x_i, Z_{ig}=1]$$

At the M step of the algorithm at the k+1 iteration we will update our parameters using the following formulas. If we use the ECME variant of the EM algorithm we do not need the $e_{1,ig}$ expectation and we continue the procedure as follows. First we update the parameters μ_g, δ_g,$\Sigma - g$:

$$\mu_g^{(k+1)} = \frac{\sum_{i=1}^{n} e_{2,ig}^{(k)} x_i - \Delta_g^{(k)} e_{3,ig}^{(k)}}{\sum_{i=1}^{n} e_{2,ig}^{(k)}}$$

$$\delta_g^{(k+1)} = (\Sigma_g^{(k)-1} x \sum_{i=1}^{n} e_{4,ig}^{(k)})^{-1} diag[\Sigma_g^{(k)-1} \sum_{i=1}^{n}(x_i - \mu_g^{(k+1)}) e_{3,ig}^{(k)'}]$$

https://doi.org/10.26219/heal.aueb.6203

$$\Sigma_g^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}[\Delta_g^{(k+1)} e_{4,ig}^{(k)'} \Delta_g^{(k+1)'} - (x_i - \mu_g^{(k+1)}) e_{3,ig}^{(k)'} \Delta_g^{(k+1)} +$$

$$+(x_i - \mu_g^{(k+1)})(x_i - \mu_g^{(k+1)})' e_{2,ig}^{(k)} -$$

$$-\Delta_g^{(k+1)} e_{3,ig}^{(k)} (x_i - \mu_g^{(k+1)})']$$

where, $\Delta_g^{(k+1)} = diag(\delta_g^{(k+1)})$ and by x we denote the elementwise product. We should mention that these formulas apply for the unrestricted case where X_o is p-dimensional vector, but it is easy to use them for the restricted case as well. Finally, since we are using the EMCE algorithm we update the parameter ν by conditionally maximize the actual log likelihood with respect to ν :

$$\sum_{i=1}^{n} logf(x_i|\mu^{(k+1)}, \delta^{(k+1)}, \Sigma^{(k+1)}, \nu)$$

As in all the previous cases we repeat the above steps until our termination criterion is satisfied. This procedure with small changes applies in all variants of the multivariate skew t distribution.

The implementation of the EM algorithm for multivariate skew t distributions is available in many packages in R. The package "EMMIXskew" uses the restricted multivariate skew t. Other packages available are the "EMMIXuskew" and the "EMMIXcskew", where the canonical fundamental skew -t distribution is fitted.

In order to understand better how mixtures of multivariate skew t distributions work in the clustering frame, we will again use the same simulated data contain- ing six different cases. This will help us evaluate this model compared to the previous ones and understand its characterists better. We will use the R package "EMMIXskew". Figure 12 shows the clusters found by this model.

We notice that using the multivariate skew t model we obtained the best solution so far. In the first five cases it gave us solutions with the actual number of clusters. It is also worth mentioning that in the sixth case we have the best fit so far since it was able to allocate the one of the two clusters regardless the peculiar shape.The Adjusted Rand Indexes are, 0.99 for the normal data, 0.78 for the t data, 0.97 and 0.92 for the cases of skew normal and skew t respectively, 0.88 for the fifth case with data from various distributions and 0.78 for the last case. The results seem in general to be better comparing to the other cases with an exception for the 5th case were the skew normal model fitted the data better.

https://doi.org/10.26219/heal.aueb.6203

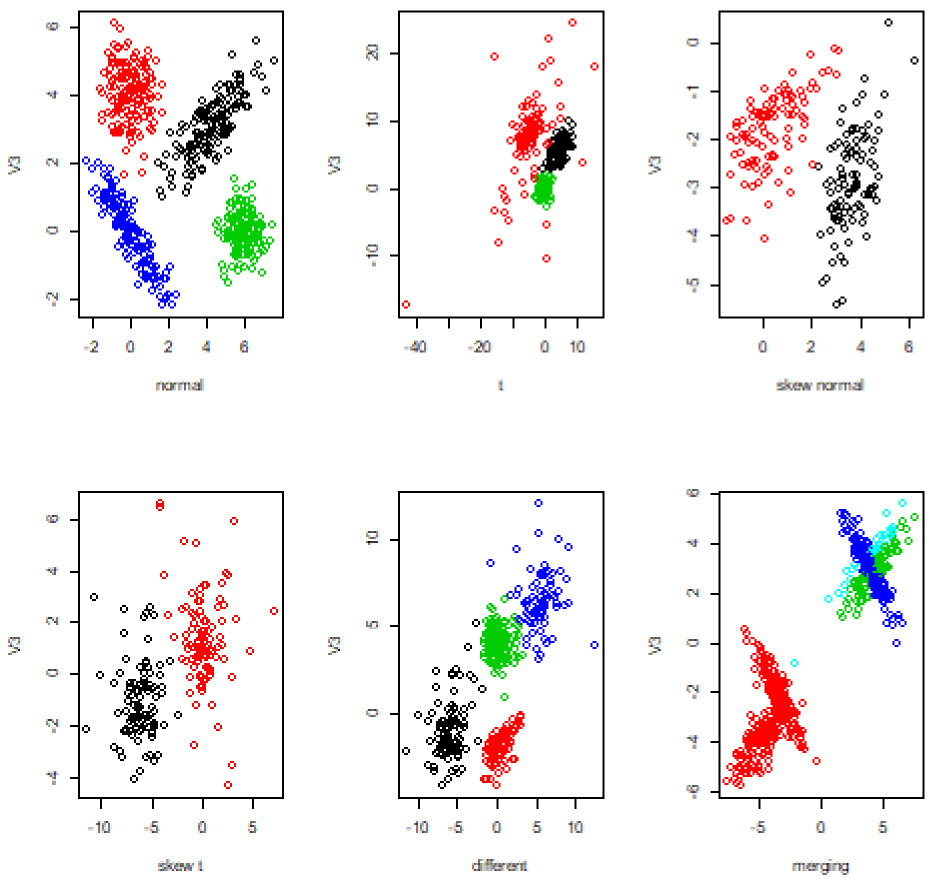


Figure 12: Fitting finite mixtures of multivariate skew t distributions in six cases of simulated data.

3.5 Merging normal components

Having studied the cases of the previous models, we saw that depending on the characteristics of the data different models have better fit each time. There are mixtures models such as the multivariate skew t, that are very flexible and able to capture at the same time heterogeneity and asymmetric characteristics. However, when the shapes of the clusters are too complicated the above mentioned models fail to allocate the correct clusters. A method to deal with this problem is to combine two or more components of a finite multivariate density in order to form a single cluster. With this approach we will be able to describe clusters of peculiar shapes that we were unable to deal with, when using the logic of the previous sections. For example in Figure 10 we can see two situations of grouped data that could not be fitted properly using the previous models.The shapes of the clusters are too strange in order to be described from these models. In order to understand better the use and the logic of this approach is that in order to define a well fitted clustering the observations within each groups should be as similar as possible and the observations between different groups should de as distanced or as different as possible. This becomes clear looking at Figure 13. Nevertheless, we should have in mind that depending on the problem, or what someone is searching using the simple mixture models described previously could be more suitable.

To begin with, the density of a finite mixture model is described as follows:

$$f(x|\Theta) = \sum_{g=1}^{G} \pi_g f_g(x|\Theta_g)$$

where Θ contains all the parameters for estimation. With $f_g(\cdot)$ we denote the density of the g-th component.

In all the previous section each component of the mixture model was considered as a single cluster. So according to the above formula we would come up with a solution containing G clusters. In the approach discussed in this section this is no longer the case. The basic idea is that the actual number of clusters is smaller (or equal) from the number of components in the mixture model. This happens because the model we use to describe the data is not flexible enough to capture all the characteristics of the data, so it uses more components. Merging mixture components suggests that in order to obtain solutions closer to the truth we should consider as one group sometimes more than one mixture component.

In order to do so, in most cases we fit the mixtures model of our choice as previously, and then we try to decide which components should be merged into one group applying some criteria that we will discuss later. Answering the question which components should be merged is not easy and most of times

https://doi.org/10.26219/heal.aueb.6203

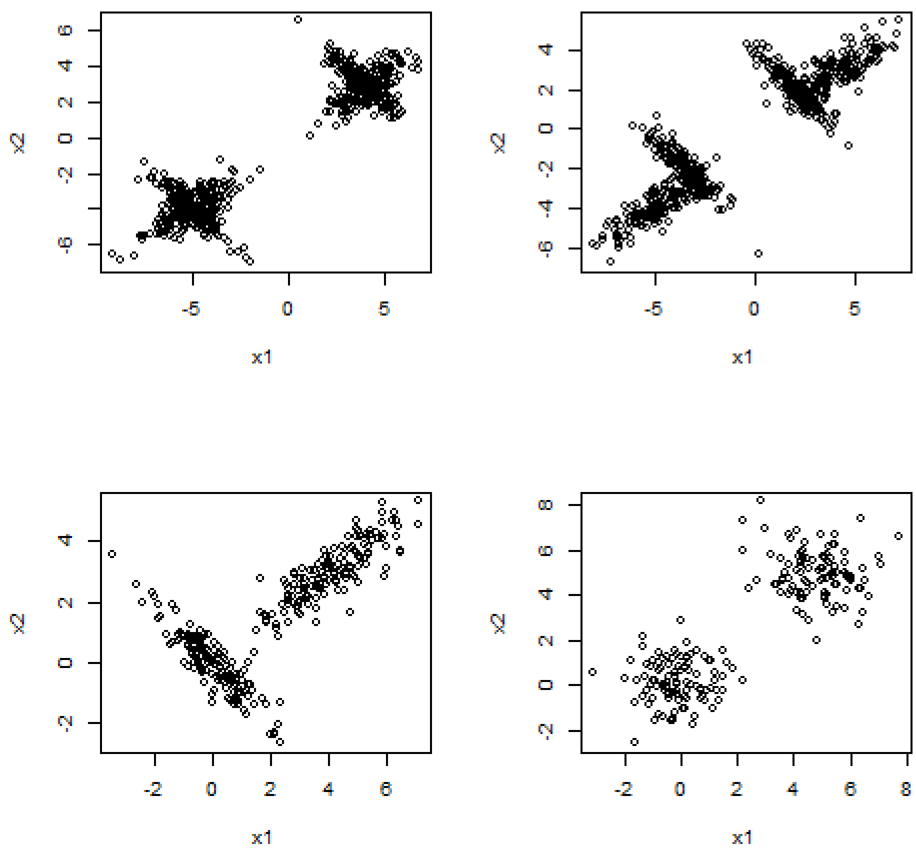


Figure 13: Examples of grouped data. In the first row we have cases where the groups should be described by combining normal mixtures components, whereas in the second line we have single component groups.

https://doi.org/10.26219/heal.aueb.6203

depends on what someone wishes to accomplice, the data etc. Figure 14 helps us understand better the fact that deciding which components should be merged some times is clear but not always. For example, for the two graphs on the right it is easier to decide that they probably should describe two different groups, whereas in the other two this could not be case. Of course before someone decides how should they proceed, it is important to take into account the hole structure of the data.

As we saw, deciding which parts should be merged is not an easy task. There have been suggested many different approaches in order to deal with this problem. Here we will refer to an approach based on an entropy criterion (Baudry et al.) and some methods based on modality and some on misclassification probabilities, for an extended review in many of them at [14].

To begin with, the number of components of the mixture models and thus, the number of clusters in the previous sections was decided using the Bayesian Information Criterion (BIC). Another criterion widely used is that of the integrated complete likelihood(ICL), which is equal to BIC penalized by the mean Entropy, given by:

$$Ent(G) = -\sum_{g=1}^{G}\sum_{i=1}^{n} t_{ig}(\theta_g) log t_{ig}(\theta_g)$$

where t_{ig} denotes the conditional probability that the i-th observation comes from the g-th component of the mixtures model. Because of its nature ICL tries to find the number of clusters instead of the number of mixtures components that BIC finds. This is the reason why, when we use it in the models discussed previously it underestimated the number of components.

Baudry et al suggested that the data should be fitted as previously choosing the number of components using BIC and then to use the Entropy to decide which components should be merged in order to find the true clusters. Analytically, the process has the following steps. First we find a model that fits the data well giving us based on the BIC the number of mixture components. Then, the idea is to construct different clustering by combining the components of the BIC solution. Each time we choose to combine the two components that minimize the Entropy of the clustering that occurred. When we cannot minimize the entropy any longer we stop and we have the desired clustering. We should mention that no matter the solution or number of clusters we end up, the fit of the data won't change. In order for this procedure to be automated the decision could be based on the ICL or the elbow graphic rule.

Other approaches as we said before are using criteria based on the modality. We should mention at this point that if we have a random variable with a density curve reaching a peak, this peak is called mode. So, if we have a density with one peak this density is said to be unimodal, if it has two bimodal etc. having this in mind it is easy to understand why the concept of modality is used in this particular problem. There exists various methods built on this concept, for example the Ridgeline method [22] or the dip test [30]. In order to understand

https://doi.org/10.26219/heal.aueb.6203

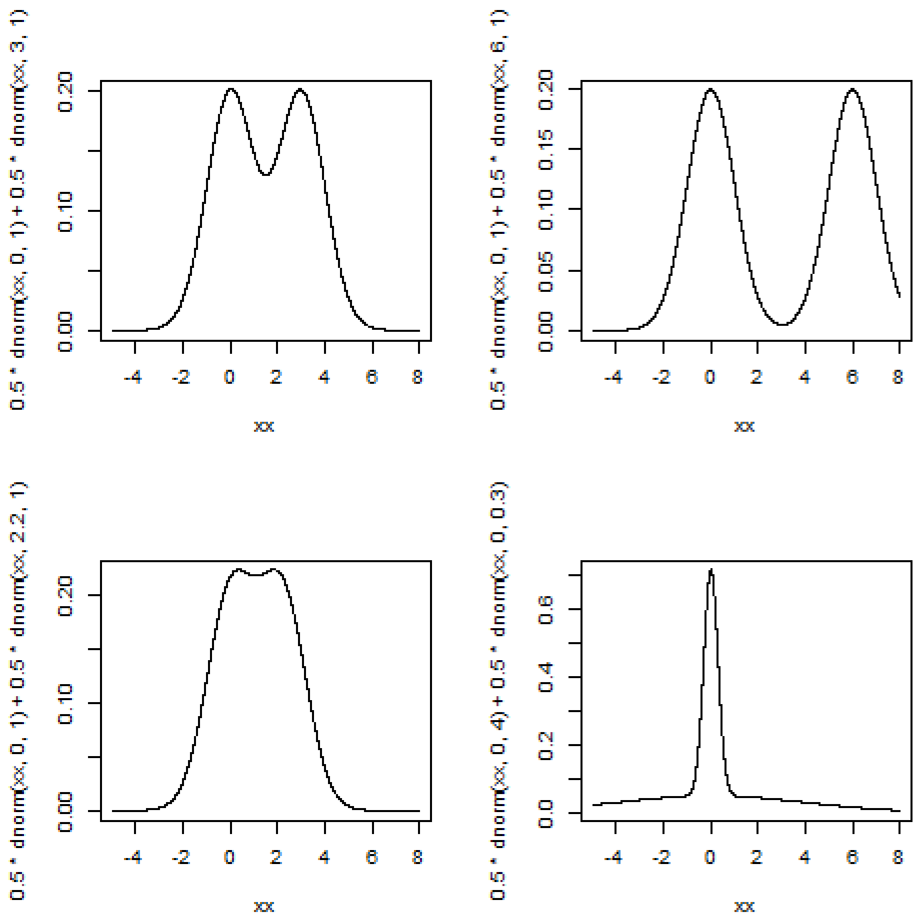


Figure 14: Various cases of mixture densities. In some situations is perhaps easier to decide whether the components should be merged.

https://doi.org/10.26219/heal.aueb.6203

better we will describe in a few words the procedure when using the Ridgeline method.

The basic idea is to choose the partition of the data in which every cluster is unimodal and any further merging of components would result in clusters with more than one peak. The following readgeline is used to find all the modes,

$$X^*(a) = [(1-a)\Sigma_1^{-1} + a\Sigma_2^{-1}]^{-1}[(1-a)\Sigma_1^{-1}\mu_1 + a\Sigma_2^{-1}\mu_2]$$

where a takes values in $[0,1]$, by Σ_1 we denote the covariance matrix of the first cluster, and by μ_1 the mean vector of the first cluster. a varies from 0 to 1 and gives a curve, this curve passes through all the modes, if someone for example plots it, it would be easy to identify the modes.

In order to find the number of the clusters, we proceed as follows. First, we fit a Gaussian mixtures model to our data and obtain the components of the model. Then for every two clusters of our current partitioning (if this is the first step each for every two components) we calculate the above ridgeline to check if the two component Gaussian mixture is unimodal or not. If none of the pairs of clusters is unimodal then we will not combine any component and we will stay with the initial partition. If the pairs are unimodal then we merge them and we continue this procedure as long as we come up with clusters that are unimodal. There are other methods using modality as we said before such as the ridgeline ratio or the dip test, the main idea remains the same but the criterion used to decided about modality is different.

Last but not least, there are approaches based on the misclassification probabilities. When using the misclassification probabilities we try to "measure" how well separated our clusters are, in this sense using this probabilities in order to built criteria for finding the optimal number of clusters leads us to an approach which is pattern based. Similarly to the modality methods, there exists here as well a great variety of tests someone could use. For example the prediction strength method [31],the directly estimated misclassification probabilities method, or the Bhattacharyya distance method which we will describe here analytically.

To start with, the Bhattacharyya distance is based on the Bayes misclassification probability and measures the distance between two distributions. Specifically,

$$d = \frac{(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)}{8} + \frac{1}{2}log(\frac{|\frac{1}{2}(\Sigma_1+\Sigma_2)|}{\sqrt{|\Sigma_1||\Sigma_2|}})$$

In order to apply this method we start by choosing a tuning constant $d^* < 1$. This constant, will determine the degree of separation between the clusters. There are many ways to determine this value, according to Henning a good value for d^* is 0.1 as it occurred by simulations from borderline distributions. Then similar to all previous methods we estimate the number of components in our data using a Gaussian mixtures model. For all pairs we compute the distance d, if $exp(-d) < d^*$ for all pairs we stop the procedure and we do not merge any of the components. If this is not the case then we merge the pair with the maximum d.

All the methods mentioned except for the entropy criterion can be implemented in R using the package "fpc".

In order to understand better how this method works and to compare it with the models from the previous sections, we will use the same simulated data that we used before with the six cases. We will use the R package "fpc" and we will choose the appropriate of the available methods depending on the case. Figure 15 demonstrates the final clusters that occurred using this method.

We notice that in the first case with the normal data none of the components was merged resulting to the true number of clusters. In the second case were the data were generated from a multivariate t distribution this method gave us a solution with 3 clusters, which is again the correct number. The correct number of clusters resulted also in the third case with the skew normal data. In the case with the skew t data the method failed to find the correct number of clusters and merged the two into one cluster. The number of clusters is also underestimated and in the fifth case. Finally, this method was able to separate correctly the data in the last case that all the previous models failed to do so. The Adjusted Rand Indexes are, 0.99 for the normal data, 0.77 for the t data, 0.96 and 0 for the skew normal and skew t cases, 0.58 for the case with the various data and 1 for the last case.

https://doi.org/10.26219/heal.aueb.6203

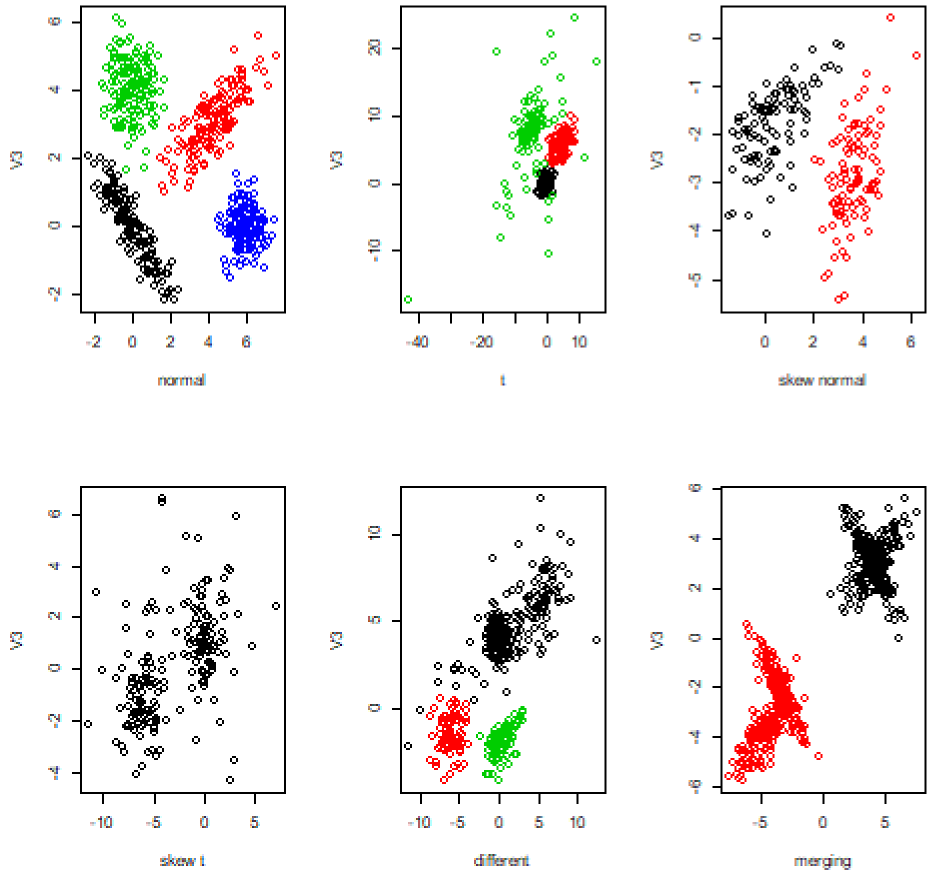


Figure 15: The resulting clustering after merging the normal mixture components.

https://doi.org/10.26219/heal.aueb.6203

3.6 Final comments

In this chapter we described analytically four different models of mixtures densities and a method that suggests that the number of clusters could be estimated more accurately if we use more than one components of a Gaussian mixture to describe a cluster. We saw that every method has advantages and disadvantages and it depends on the problem and the data to choose the best suited each time.

Using the Gaussian mixtures model, resulted in elliptical shaped clusters, making it difficult to fit data with non normal characteristics such as asymmetry and heterogeneity, so in most cases it overestimated the number of clusters. Skew normal mixtures were able to capture better skewed data but again failed in situations were a "fat-tail" behavior was exhibited. On the other hand t mixtures were able to fit data with the opposite characteristics. The Skew t mixtures model proved to be the most flexible capturing both of the above mentioned characteristics. However, in situations with strange shaped clusters all methods failed except for the combining normal components method. The following matrix summarizes the ARI results for the simulated data from all methods.

ARI					
Data	Gaussian mix.	t mix.	Skew nor.	skew t mix.	merging
normal data	0.99	0.99	0.52	0.99	0.99
t data	0.75	0.85	0.30	0.78	0.77
skew nor.data	0.72	0.96	0.92	0.97	0.96
skew t data	0.71	0.88	0.68	0.92	0
various	0.91	0.88	0.95	0.88	0.58
odd shaped	0.5	0.5	0.5	0.71	1

Table 2:The Adjusted Rand Indexes for all models and data.

We notice that skew t and t multivariate mixture models had a better overall performance, especially the skew t model, where the lowest value of ARI was 0.71. On the other hand the merging method gave both very good and very bad results with ARI from 0 to 1. The performance of the skew t model was again in the same frame varying from a very poor fitting (ARI=0.3) to very good results(ARI=0.95).

 https://doi.org/10.26219/heal.aueb.6203

4 Application

In the previous sections we described four different mixtures models for clustering and an alternative approach suggesting that we should combine more than one normal components sometimes in order to estimate the number of clusters. We evaluated the performance of these methods in simulated data representing six different scenarios. We were able that way to understand better how each model behaves under certain situations, as well as a rough idea regarding under which circomstances, each model is more appropriate.

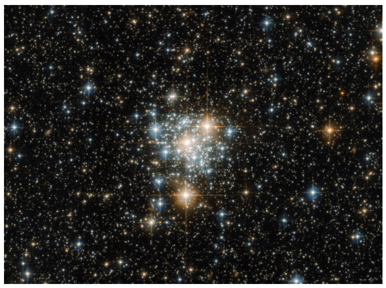
These six simulated cases however, are vary restricted. Real data on the other hand can be proven way more demanding and excibite various characteristics not described above. Thus, it is useful and interesting to evaluate the performance of these methods, make comparisons between them and see how flexible they are in real data situations.

As we said in the Introduction clustering has many applications in a huge variety of fields. Astronomy is one of these fields, and deals with many interesting problems such as the clustering of galaxies, stars , and other objects. Here we will deal with the problem of star clustering using real data. In section 4.1 we will describe analytically the problem and the data that we will use. We will divide our data in two cases one with a small number of groups and one with a large number, and we will represent the results in sections 4.2 and 4.3.

4.1 The problem

Before we continue with the application and the technical details, it would be useful and interesting to describe the nature of the problem. As we said earlier the problem is to identify young star clusters. In our Galaxy (and general in galaxies) there are regions where stars are being formed. During this procedure stars form pairs and groups, these groups are known as open clusters and they consist of some hundreds of particularly young stars most of the times. These regions are found within the spiral arms of galaxies.(Figure 16)

The study of such clusters is important because they are connected to star



(a) The NGC 299 open cluster

(b) The LH95 star forming region

Figure 16: Open clusters of young stars.These images are from the hubble telescope.

formation, a procedure that it is not yet fully understood. Stellar evolution theories are also based on the observation of star clusters. Furthermore, clusters are used in order to determine the distance scale of our universe. These are some of the fields in Astronomy that the understanding of star clusters is useful. One of the ways to collect such data, is X-ray emission, as a way to identify young stars.

The data that we will use here are from the MYStIX survey (Feigelson et al.,2013). This project contains 31745 young stars in 17 star forming regions[18]. The dataset has many variables but here we will use only two of them. These variables are the two angular coordinates of a star position in the sky, namely declination and right ascension. specifically, these are astronomical coordinates used to describe an object's position in the celestial sphere.

In this application we will use only 2 of the regions of the dataset in order to find the subclusters. We will choose the region of Orion and Carina, because of the different characteristics they have. The first one contains fewer observations compared to the other one and they have different morphological characteristics. Figure shows how the data of these two subsets look like. Also from this Figure we have an idea that the Orion region includes a smaller amount of subclusters of stars compared to Carina subset. These characteristics will help us evaluate our models in different situations.

We will examine each subset seperatly by fitting all the previously mentioned models and approaches. We will try to evaluate the quality of the clustering each time. in order to do so, we will use the Wilk's Λ measure. This statistic is used in MANOVA for testing the equality between vectors of means, it measures the within variability with respect to the total variability. It is calculated as

https://doi.org/10.26219/heal.aueb.6203

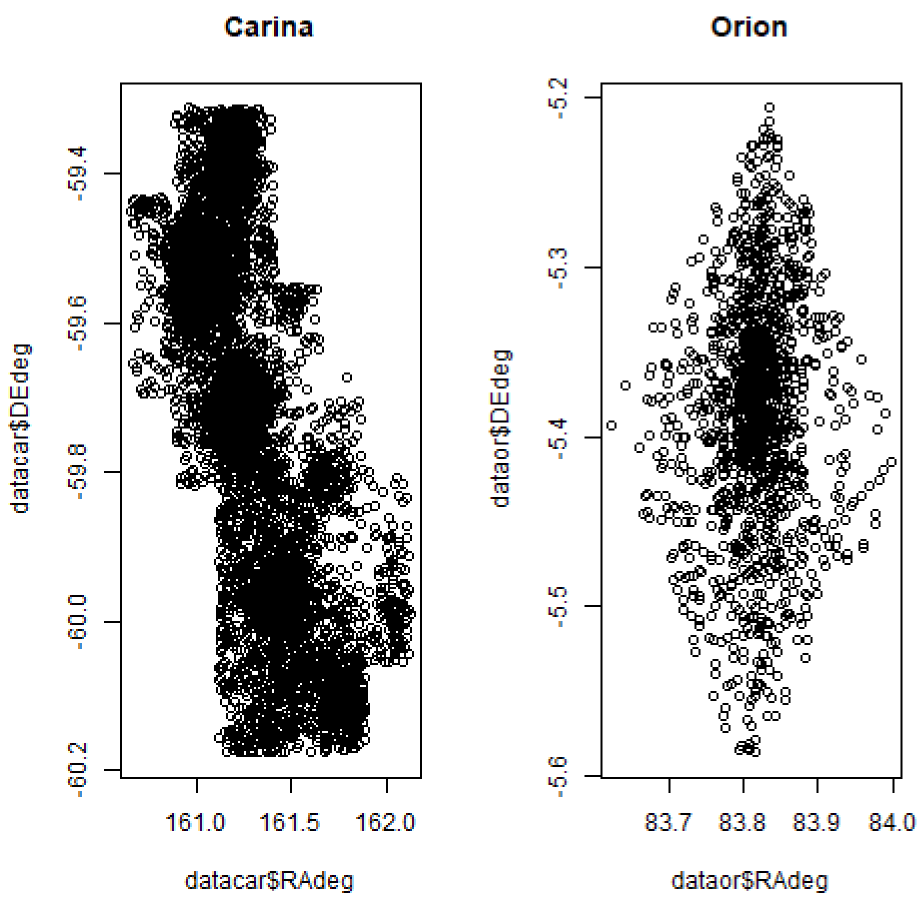


Figure 17: The Carina and Orion regions of star clusters.

https://doi.org/10.26219/heal.aueb.6203

follows:

$$\Lambda = \frac{|W|}{|W + B|}$$

where B is the between groups variability and W the within group variability. Given the nature of the data this statistic is considered suitable in order to provide us with a measure of how good a cluster is. Small values indicate differences between the groups and thus a good partition of the data.

Then we will compare the models and using ARI we will try to assess how similar the solutions are. At this point it is important to mention that we will use the same packages in R tha we used in Section 3. Also, because the number of observations combined with the large number of components in the mixture models (especially for the Carina region), makes the results unstable we fit various times from different starting points and choose the best solution among them.

4.2 1st case

As we said previously, we chose two regions to work with, one with fewer observations and potentially smaller number of clusters from the other. In this section we will analyze the data from the Orion region. This data subset contains 1524 observations (or stars). We will fit all the previously mentioned models and the merging method, and we will discuss analytically the solution each of them gave us. Then we will try to evaluate the clustering that occurred from each model using the Wilk's Λ statistic. Finally, we will compare the solutions and we will try to assess how similar the results are one with another using the Adjusted Rand Index.

To begin with, we start by fitting the multivariate Gaussian mixture model. As we described in the previous sections we choose between 14 different models depending on the elements from the eigenvalue decomposition of the covariance matrix. The optimal solution for this particular situation was a clustering consisting of 4 groups that have different volume, shape and their orientation is accordingly to the coordinate axes. Figure 16 illustrates the clusters that we came up. We notice that the groups are very different one with another, the solution consists of groups very wide and some very dense.

The exact clustering can be found in table 3 were we can see how many observations each cluster has. We see that the number of observations varies a

https://doi.org/10.26219/heal.aueb.6203

lot between the groups with one group for example having 564 observations and another 5 times fewer.

Clustering Table			
Group1	Group2	Group3	Group4
564	122	397	441

Table 3: Clustering from multivariate Gaussian mixture.

The Wilk's Λ is equal to 0.83. This is a relatively high score indicating a poor solution in terms of clustering, since as we said before, we need values close to zero, in order for our clustering to be good.

Next, we will fit to our data the multivariate t mixture model. Here, the number of models that we can choose between is 28. The results in this situation indicated a solution with six clusters, with different volume, same shape and orientation accordingly to the coordinate axes. Figure 19 illustrates the final clustering in detail. we notice that the groups seem quite different from those with the Gaussian model. We see that here as well there are groups very dense and others quite scattered. It is also clear from the graph that the number of observations varies a lot between the groups. More details on the exact clustering can be found in table 4, which shows the number of observations in each cluster. We see that difference is quite intense since we have a group containing 20 observations and a group with 591.

Clustering Table					
Group1	Group2	Group3	Group4	Group5	Group6
95	20	395	591	379	46

Table 4:Clustering from multivariate t mixtures(Orion region).

The value of Wilk's Λ in this situation is equal to 0.64. It is again a quite large number however it is quite smaller from the corresponding value at the Gaussian mixture model.

Moving on, we will fit the multivariate Skew Normal mixture model to our dataset. At this point we should mention that the restricted distribution will be fitted and we will choose between 5 different models depending on the form of the covariance matrix. The results, indicate a solution with 6 clusters. The resulting clustering seems quite different from the solution of the t mixtures which was again with six clusters. Figure 18 demonstrates the clustering occurred from this model in detail. We notice that the clusters seem

https://doi.org/10.26219/heal.aueb.6203

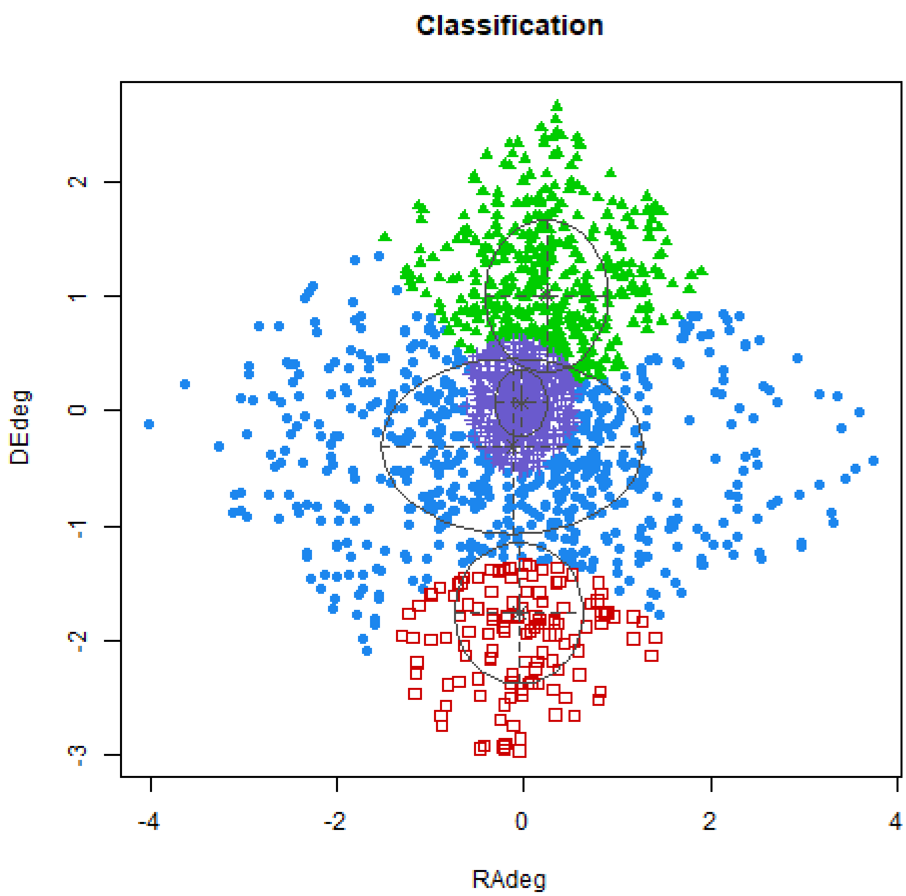


Figure 18: The clustering with the multivariate Gaussian mixture model. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

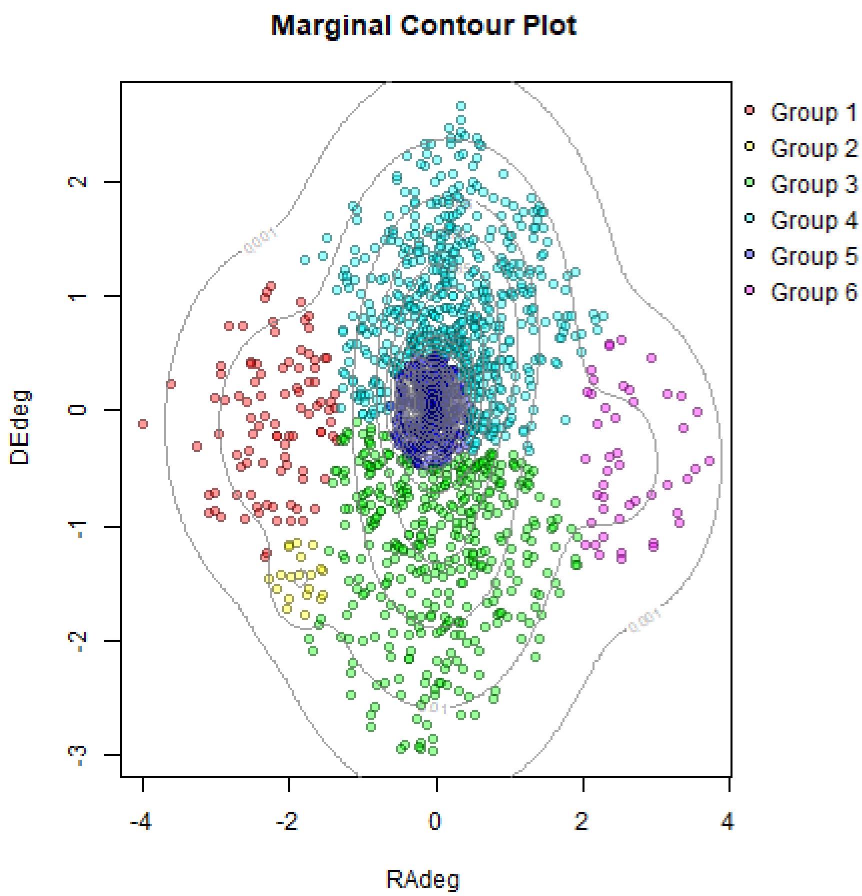


Figure 19: The clustering with the multivariate t mixture model. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

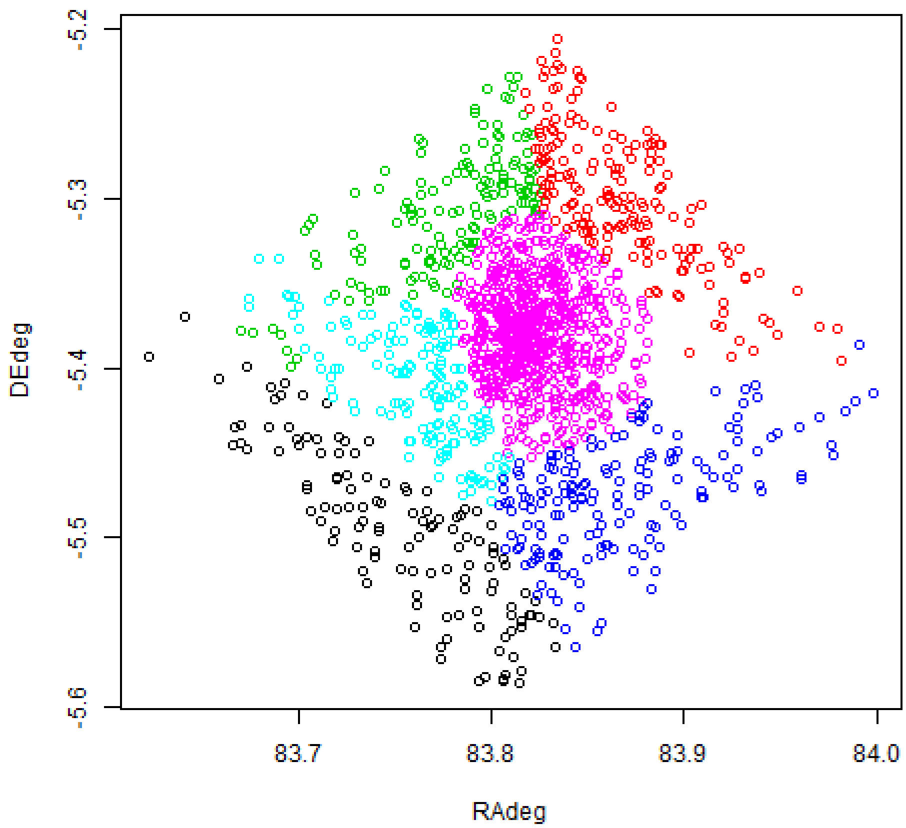


Figure 20: The clustering with the multivariate Skew Normal mixture model. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

very similar one with another except for the cluster in the center. They also look similar in size. Table 5 illustrates the clustering analytically with the exact number of observations in each group. We notice that in 5 out of the 6 groups the number of observations is similar except for the sixth group that contains the larger part of the observations.

Clustering Table					
Group1	Group2	Group3	Group4	Group5	Group6
122	182	145	166	162	747

Table 5:Clustering from multivariate Skew Normal mixtures (Orion region).

The Wilk's Λ for this model is equal to 0.99. The value is very large since it is almost equal to 1. This indicates that this is not a good clustering. This value is also the larger one so far.

Next, we will fit the multivariate Skew t mixture model. Similarly to the skew normal case only the restricted distribution is fitted and we again choose between 5 different types of scale matrix. The results in this situation indicate a solution with 4 clusters, similar to the solution with the normal mixtures. Figure 19 illustrates the clustering from this model. We notice that we have again clusters more dense than the others, and also clusters different in size and shape.

More details regarding the exact clustering can be found in table 6. We see that the smaller cluster contains 159 observations and the larger one 653. The Wilk's Λ here is equal to 0.73, it is considered again a relatively large value, but it is the second lowest comparing with the other models.

Clustering Table			
Group1	Group2	Group3	Group4
429	653	159	283

Table 6:Clustering from multivariate Skew t mixtures (Orion region).

Finally, we tried to apply the method described in section 3.5 that suggested that the clusters should emerge by combining the normal mixture components we found before. We used various different test as described in the relevant section and all indicated a solution with one cluster.

Having seen analytically all the solutions provided with the above mentioned models, it would be of interest at this point to compare them with each other.

https://doi.org/10.26219/heal.aueb.6203

Overall, the Gaussian mixture model and the skew t mixture model gave us a solution with 4 clusters. The t mixture model and the skew normal indicated results with 6 groups. The Wilk's Λ was relatively hight in all situations, in some cases it was almost 1. The smallest value was noticed for the t mixture model indicating it as the best amongst the others.

The solutions that came from the different models were quite different one with the other, as we saw previously. Table 7 illustrates the Adjusted rand Index between the possible clusterings, in an attempt to evaluate how similar the solutions are.As expected the values of ARI are relatively low, indicating that we have different clustering depending on the model. This is due to the fact that as we saw the fitting was not good enough, this has also to do with the nature of the data.

ARI					
model	Gaussian mix.	t mix.	Skew nor.	skew t mix.	merging
Gaussian mix.	1	-	-	-	-
mix	0.38	1	-	-	-
Skew nor.	0.25	0.24	1	-	-
skew t mix	0.30	0.31	0.27	1	-
merging	-	-	-	-	1

Table 7:ARI-Comparisons between models, if there were 6 clusters for all.

https://doi.org/10.26219/heal.aueb.6203

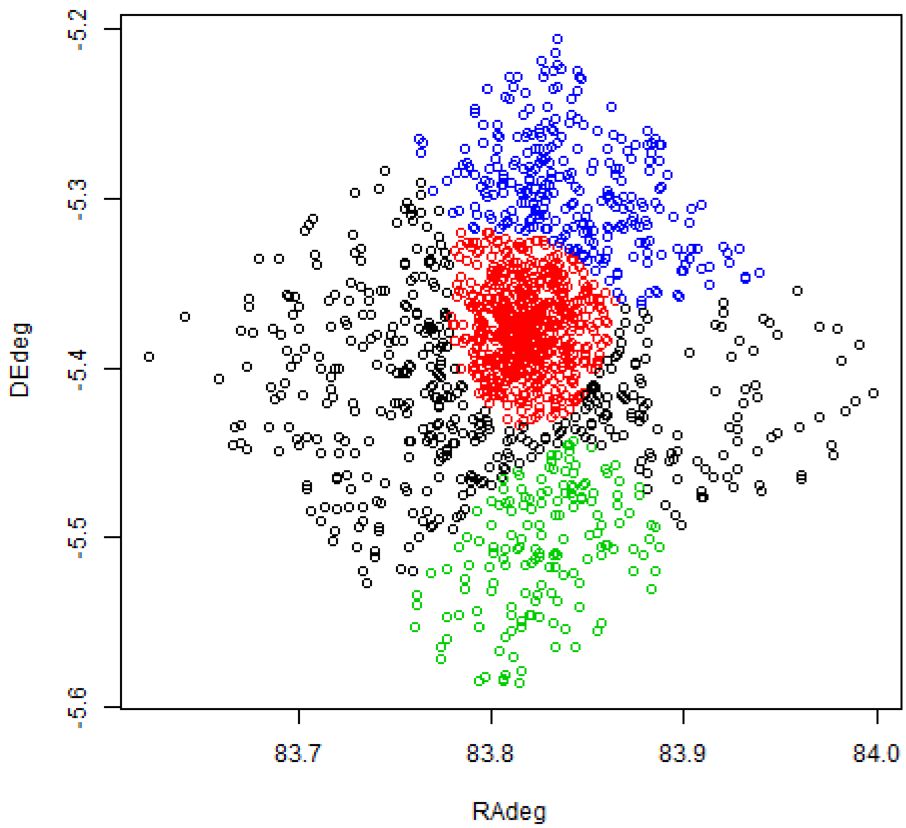


Figure 21: The clustering with the multivariate Skew t mixture model. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

4.3 2nd case

As we mentioned previously we will examine two subsets of our dataset. We have already seen and described analytically the smaller one with observations from the Orion region, and now we will follow the same procedure for the subset from the Carina region. This dataset contains 7334 observations and has also a larger number of groups. Thus, it would be interesting to see how these models would fit a dataset like this.

To start with, we will fit all the models and approaches described in the previous sections. We will evaluate the performance of each model using again the Wilk's Λ based on these we will decide which model provides us with the best clustering. Furthermore, it would be of interest to compare the solutions and see how similar the results are.

We will first start by fitting the multivariate Gaussian mixture model. As we said previously using this model we can choose between 14 different models. The results indicated a solution with 18 clusters with varying volume and shape, and oriented according to the coordinate axes. Figure 22 illustrates the resulting clustering with details. We notice that the groups differ a lot in size and density. Due to the nature of the data the groups are really close one to another making the problem of group identification difficult, however we can see that there seems to be groups more distinct from the rest.

More details on the clustering can be found in table 8 which illustrates the exact number of observations in each group. We notice that there are groups with as many observations as 1257 while on the other hand there are groups with almost 40 observations.The Wilk's Λ for this clustering is equal to 0.97, really close to 1, indicating a very poor fitting of the data in terms of clustering.

Clustering Table					
Group1	38	Group7	957	Group13	137
Group2	236	Group8	600	Group14	1257
Group3	304	Group9	505	Group15	332
Group4	131	Group10	885	Group16	86
Group5	882	Group11	135	Group17	241
Group6	464	Group12	79	Group18	65

Table 8:Clustering from the multivariate Gaussian model(Carina region).

Next, we will fit to the data the multivariate t mixtures model. The results

https://doi.org/10.26219/heal.aueb.6203

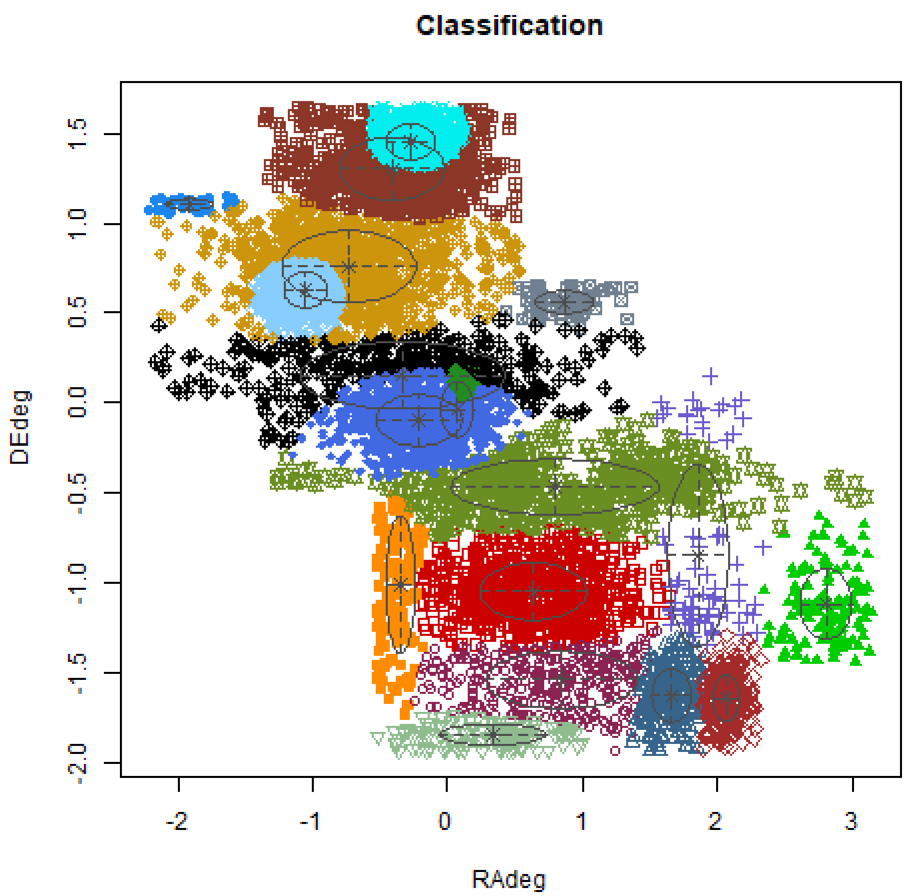


Figure 22: The clustering with the multivariate Gaussian mixture model for the Carina region. Different colors correspond to different clusters.

 https://doi.org/10.26219/heal.aueb.6203

indicated a clustering consisting of 17 groups with different volume and shape, and orientation accordingly to the coordinate axes, similar to the Gaussian model. Figure 21 illustrates analytically the clustering. We notice that the clusters have again the same characteristics as previously. In Table 9 we can see the exact number of observations in each cluster. The size of the clusters is again similar to the previous clustering with very large and very small groups. The Wilk's Λ for this clustering is 0.32, a relatively low value especially compared to the value for the Gaussian solution. As we mentioned before we need small values because they indicate a good clustering.

Clustering Table					
Group1	37	Group7	138	Group13	789
Group2	1262	Group8	270	Group14	130
Group3	268	Group9	925	Group15	455
Group4	795	Group10	316	Group16	99
Group5	1124	Group11	65	Group17	138
Group6	131	Group12	392		

Table 9:Clustering from the multivariate t model(Carina region).

Next, we will fit the multivariate Skew normal mixture model. The restricted distribution is only fitted and we can choose between five different types for the scale matrix. The results, indicate a solution with 13 clusters. Figure 22 illustrates in detail the clustering that occurred. this model gave us a clustering with the number og groups reduced by five compared to the gaussian model and by four compared to the t mixture model. Table 10 shows analytically the number of observations in each cluster. The Wilk's Λ for this model is equal to 0.43. This value is again quite smaller from the equivalent occurred from the multivariate Gaussian mixtures.

Clustering Table					
Group1	1035	Group6	58	Group10	145
Group2	152	Group7	190	Group11	1031
Group3	118	Group8	207	Group12	850
Group4	1636	Group9	895	Group13	470
Group5	547				

Table 10:Clustering from the multivariate Skew Normal model(Carina region).

https://doi.org/10.26219/heal.aueb.6203

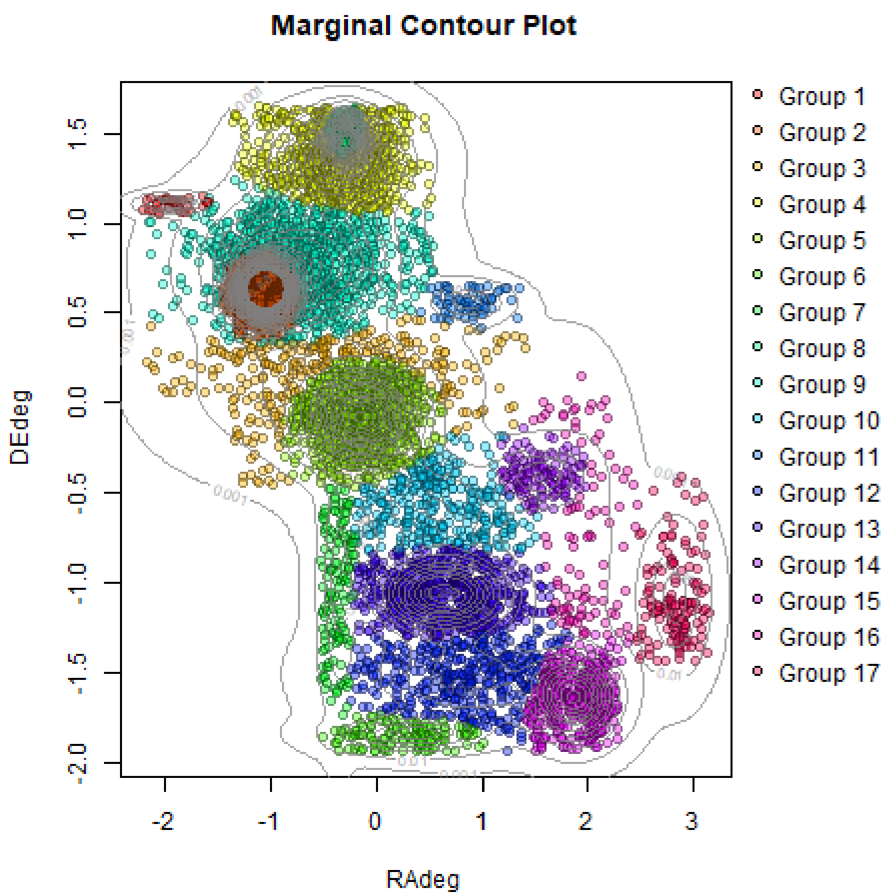


Figure 23: The clustering with the multivariate t mixture model for the Carina region. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

Next is the multivariate skew t mixture model. The results here indicate a solution with 10 clusters. this number is the smallest compared to the previous ones. Figure 23 shows the clustering in detail. We notice that the outcome seems more different from the others. More information regarding the size of each cluster can be found in Table 11 . The Wilk's Λ here is 0.51, it is a rather large value but it is still better than the solution from the Gaussian mixtures.

Clustering Table			
Group1	327	Group6	340
Group2	972	Group7	176
Group3	2222	Group8	1187
Group4	899	Group9	310
Group5	363	Group10	538

Table 11:Clustering from the multivariate Skew t model(Carina Region.)

Finally, we will apply the merging method. We tried all the tests mentioned above. All results indicated to the solution of two clusters. However, this is not very representative of the truth. The reason this method provides us with poor results is perhaps due to the nature of the data. Being this way either modality or misclassification methods would fail to distinct between the clusters.

Overall, we saw that each model gave us a different number of clusters. With the number varying from 18 groups (Gaussian model) to 10 groups (skew t model). What is more from the Wilks Λ statistic it was clear that the best clustering resulted from the t model followed by the skew normal. Table 12 shows The Adjusted Rand Indexes for the clusterings in order to examine how similar the solutions are. We notice that in most cases the results agree to an extend especially in the case of Gaussian and mixtures. The fact that all the solutions are not that different is a good sign that our clustering is not bad, and that the data have indeed a structure.

ARI				
model	Gaussian mix.	t mix.	Skew nor.	skew t mix.
Gaussian mix.	1	-	-	-
mix	0.74	1	-	-
Skew nor.	0.68	0.70	1	-
skew t mix	0.61	0.64	0.77	1

Table 12:ARI-Comparisons between models, if all models had 17 clusters.

https://doi.org/10.26219/heal.aueb.6203

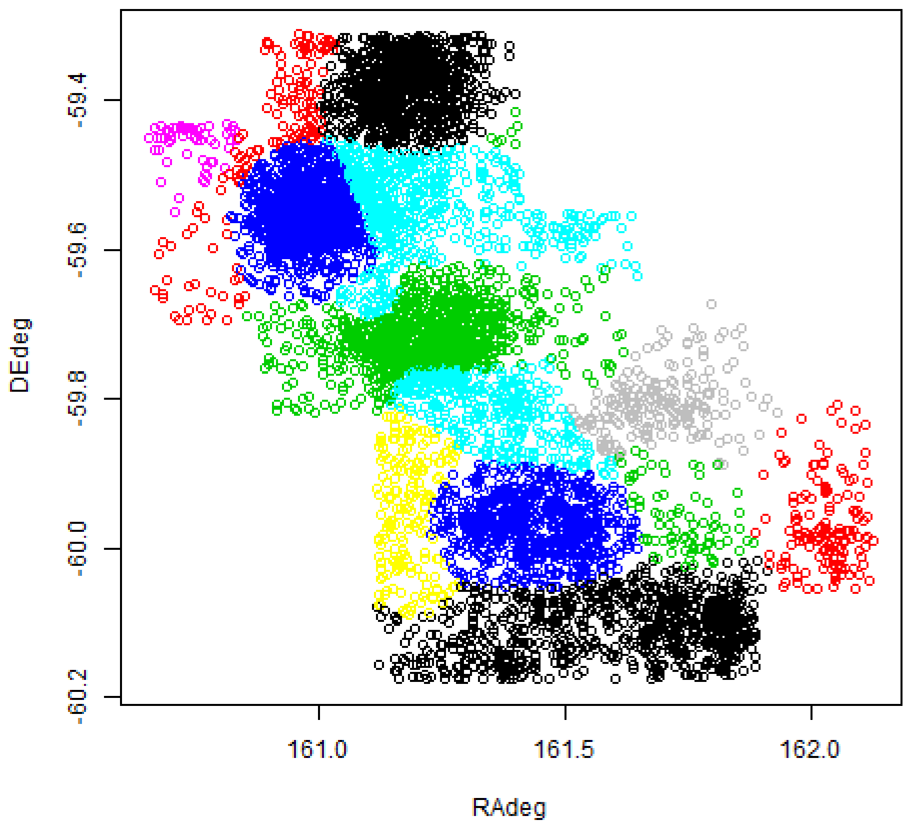


Figure 24: The clustering with the multivariate Skew Normal mixture model for the Carina region. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

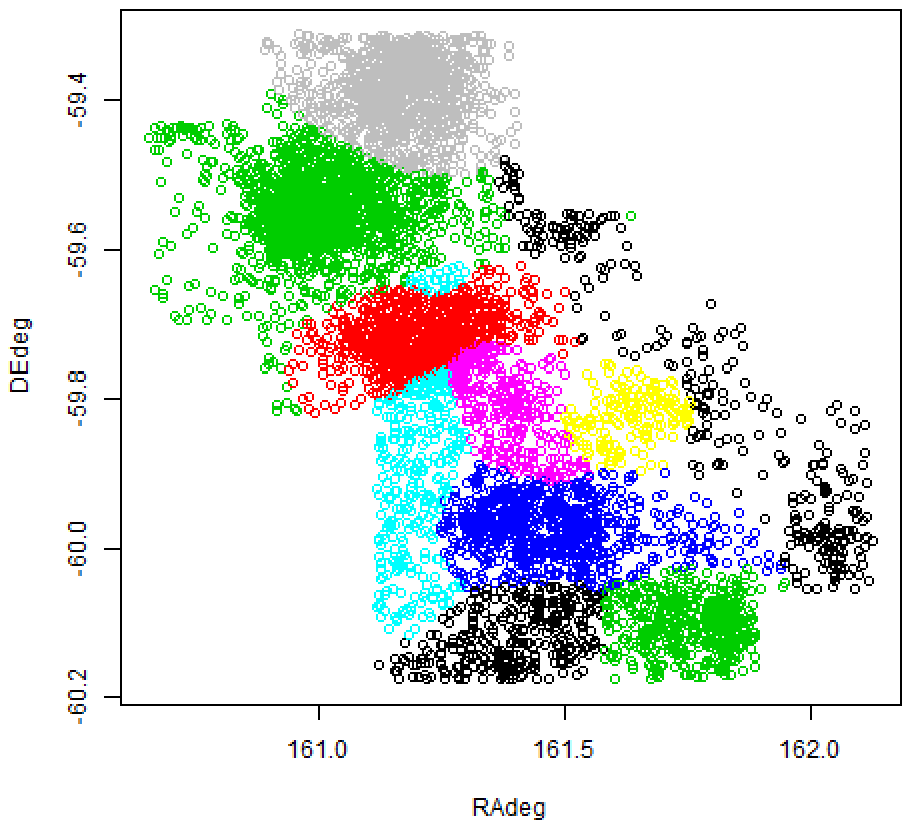


Figure 25: The clustering with the multivariate Skew t mixture model for the Carina region. Different colors correspond to different clusters.

https://doi.org/10.26219/heal.aueb.6203

5 Conclusion

As we saw model based clustering is a very useful technique with many advantages compared to other methods. It is widely used in many areas and has gained a lot of attention over the last decades. There exists a huge variety of models and methods for model based clustering, each with specific advantages depending on the problem we are dealing with. We saw and examine in detail five different models and methods. We started with the multivariate Normal mixtures model and then proceeded to multivariate t, skew normal, skew t and finally the combining normal components method. We applied these methods to six cases of simulated data and we saw that depending on the dataset each time different method performed better.

In situations with more heterogeneous data we saw that multivatie t and multivariate skew t distributions performed better. When we were dealing with data with asymmetric characteristics the skew models provided us with better solutions. Generally the multivariate skew t model seemed to be more flexible compared to the others since it adapted in most situations.However in situations where the data were forming odd shaped clusters most of these models failed, and only the merging method was able to identify correctly the clusters.

We then applied all these techniques to a real dataset with observations of young stars. Clusters of young stars is an interesting problem both in terms of Astronomy and in terms of Statistics. We examine two different subsets of data with different characteristics. The first one was a dataset with observations from the Orion region. This subset contained fewer observations compared to se second one, had a relatively small number of clusters, but the groups weren't well seperated. On the other hand the data from the Carina region contained a large number of observations and also had many clusters that were however, better seperated.

In both cases, the multivariate t mixtures model provided us with the best clustering, and normal the worst. This result is logical given that t distribution deals with the "fat-tail" problem a characteristic that was present in these data. The Skew models did not gave us the better solution because as we saw the data did not expedited intense asymmetric behavior. The merging method did not proved useful in this problem, because due to the nature of the data it tended to merge them all in one or two groups. Finally we saw that in the Carina case the methods seem to agree in a more significant level, than in the orion region. Seeing all these, it is clear that we should choose which method to use depending on the problem and the data, since different characteristics are treated differently from each model.

As we also saw, there are still a lot more possible models or techniques that someone could apply. Regarding this specific problem it would be interesting to

https://doi.org/10.26219/heal.aueb.6203

see how different models, like the normal inverse gaussian, the shifted asymmetric Laplace or the use of copulas, would have fitted these data, and also examine how similar the results would be. Other criteria of clustering evaluation could have also been used in order to assess our clustering and compare the possible solutions.

References

[1] Jeffrey L Andrews and Paul D Mcnicholas. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, 22(5):1021–1029, 2012.

[2] Jeffrey L Andrews, Paul D McNicholas, and Sanjeena Subedi. Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics & Data Analysis*, 55(1):520–529, 2011.

[3] JL Andrews and PD McNicholas. teigen: Model-based clustering and classification with the multivariate t-distribution. *R package version*, 1:24–46, 2012.

[4] Reinaldo B Arellano-Valle and Marc G Genton. On fundamental skew distributions. *Journal of Multivariate Analysis*, 96(1):93–116, 2005.

[5] Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389, 2003.

[6] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

[7] Jean-Patrick Baudry, Adrian E Raftery, Gilles Celeux, Kenneth Lo, and Raphael Gottardo. Combining mixture components for clustering. *Journal of computational and graphical statistics*, 19(2):332–353, 2010.

[8] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.

[9] Márcia D Branco and Dipak K Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

[10] Celso Rômulo Barbosa Cabral, VíCtor Hugo Lachos, and Marcos O Prates. Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis*, 56(1):126–142, 2012.

[11] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[12] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[13] Francesca Greselin and Salvatore Ingrassia. Constrained monotone em algorithms for mixtures of multivariate t distributions. *Statistics and computing*, 20(1):9–22, 2010.

[14] Christian Hennig. Methods for merging gaussian mixture components. *Advances in data analysis and classification*, 4(1):3–34, 2010.

[15] Dimitris Karlis and Anais Santourian. Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, 19(1):73–83, 2009.

[16] Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590, 2003.

[17] Ioannis Kosmidis and Dimitris Karlis. Model-based clustering using copulas with applications. *Statistics and computing*, 26(5):1079–1099, 2016.

[18] Michael A Kuhn, Eric D Feigelson, Konstantin V Getman, Adrian J Baddeley, Patrick S Broos, Alison Sills, Matthew R Bate, Matthew S Povich, Kevin L Luhman, Heather A Busk, et al. The spatial structure of young stellar clusters. i. subclusters. *The Astrophysical Journal*, 787(2):107, 2014.

[19] Sharon Lee and Geoffrey J McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202, 2014.

[20] Sharon X Lee and Geoffrey J Mclachlan. On mixtures of skew normal and skew

t

-distributions. *Advances in Data Analysis and Classification*, 7(3):241–266, 2013.

[21] Sharon X Lee and Geoffrey J McLachlan. Emmixcskew: an r package for the fitting of a mixture of canonical fundamental skew t-distributions. *arXiv preprint arXiv:1509.02069*, 2015.

[22] Jia Li, Surajit Ray, and Bruce G Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(Aug):1687–1723, 2007.

[23] Tsung I Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100(2):257–265, 2009.

[24] Tsung I Lin, Jack C Lee, and Wan J Hsieh. Robust mixture modeling using the skew t distribution. *Statistics and computing*, 17(2):81–92, 2007.

https://doi.org/10.26219/heal.aueb.6203

[25] Chuanhai Liu and Donald B Rubin. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence. *Biometrika*, 81(4):633–648, 1994.

[26] Paul D McNicholas. Model-based clustering. *Journal of Classification*, 33(3):331–373, 2016.

[27] Paul David Mcnicholas and Thomas Brendan Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

[28] Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[29] Katherine Morris and Paul D McNicholas. Dimension reduction for model-based clustering via mixtures of shifted asymmetric laplace distributions. *Statistics & Probability Letters*, 83(9):2088–2093, 2013.

[30] Jeremy Tantrum, Alejandro Murua, and Werner Stuetzle. Assessment and pruning of hierarchical model based clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 197–205. ACM, 2003.

[31] Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[32] I Vrbik and PD McNicholas. Analytic calculations for the em algorithm for multivariate skew-t mixture models. *Statistics & Probability Letters*, 82(6):1169–1174, 2012.