

**ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS

**SCHOOL OF INFORMATION SCIENCES  
& TECHNOLOGY**

**DEPARTMENT OF STATISTICS**

**POSTGRADUATE PROGRAM**

**Model Based Clustering For High Dimensional Data**

By

Vassiliki Emilianos Paizi

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

June 2016





**ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS

**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ  
ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ**

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ**

**ΟΜΑΔΟΠΟΙΗΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ  
ΓΙΑ ΔΕΔΟΜΕΝΑ ΜΕΓΑΛΩΝ ΔΙΑΣΤΑΣΕΩΝ**

**Βασιλική Αιμιλιανού Παΐζη**

**ΔΙΑΤΡΙΒΗ**

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

Αθήνα  
Ιούνιος 2016



## **DEDICATION**

To my parents  
and to my teachers...



## ACKNOWLEDGEMENTS

Firstly, I would like to thank my thesis advisor Prof. Karlis Dimitris of Department of Statistics at Athens University of Economics and business. The door to his office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

Last but not the least, I would like to thank my family: my parents and my brothers for supporting me spiritually throughout writing this thesis and my life in general.

Paizi Vassiliki





## VITA

I was born in Athens in 1989. After completing my studies at Petroupolis High School, I went on to the National and Kapodistrian University of Athens, where I studied mathematics and received my Bachelor degree in June 2013. During my studies, I realized my interest in science statistics. As a result I continued my studies in the department of Statistics at Athens University of Economics and Business, in September 2014, receiving a master of Statistics in June 2016.

During my studies, I used to work as a mathematician by preparing students for panhellenic exams.





## ABSTRACT

Vassiliki Paizi

### **Model Based Clustering For Large Data**

June 2016

The rapid increase in the size of data sets makes clustering all the more important to capture and summarize the information; at the same time clustering is more difficult to be accomplished. If model-based clustering is applied directly to a large data set, it is too slow for practical application. A simple and common approach is to first cluster a random sample of moderate size, and then use this clustering model in this way to classify the remainder of the objects.

During the last years, model Based clustering has gained considerable interest. In this thesis we will investigate the problem of applying the methodology to data sets with large  $n$  and large  $p$ . The approach of using mixture of Factor Analyzers will be developed for certain competitive models. We provide an extensive review of the method and we apply the methodology to real data set providing some insight to the approach.





## ΠΕΡΙΛΗΨΗ

Βασιλική Παΐζη

### ΟΜΑΔΟΠΟΙΗΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ ΓΙΑ ΜΕΓΑΛΗ ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Ιούνιος 2016

Η γρήγορη αύξηση του μεγέθους του συνόλου δεδομένων, καθιστά τη διαδικασία ομαδοποίησης όλο και πιο σημαντική, καθώς συλλαμβάνει και συνοψίζει την πληροφορία, την ίδια στιγμή που η διαδικασία αυτή γίνεται όλο και πιο δύσκολη. Αν η ομαδοποίηση με τη χρήση μοντέλων εφαρμόζεται απευθείας σε μεγάλο σετ δεδομένων, η εφαρμογή αυτή στην πράξη μπορεί να είναι χρονοβόρα. Μια απλή και κοινή προσέγγιση είναι να ομαδοποιήσουμε πρώτα ένα τυχαίο δείγμα μέτριου μεγέθους κι έπειτα να χρησιμοποιήσουμε αυτό το μοντέλο για να κάνουμε την ανάλυσή μας.

Η ομαδοποίηση με τη χρήση μοντέλων για μεγάλη βάση δεδομένων έχει αποκτήσει ιδιαίτερο ενδιαφέρον τα τελευταία χρόνια. Στην παρούσα εργασία θα διερευνηθεί το πρόβλημα της εφαρμογής σε μεγάλες βάσεις δεδομένων  $n$  με μεγάλη διάσταση  $p$ . Θα αναπτυχθεί η προσέγγιση με τη χρήση mixture of Factor Analyzers για ορισμένα ανταγωνιστικά μοντέλα. Παρουσιάζουμε τα κύρια σημεία της μεθόδου, εφαρμόζοντας τη μεθοδολογία σε πραγματικά δεδομένα ώστε να μπορέσουμε να κατανοήσουμε την παραπάνω προσέγγιση.







## TABLE OF CONTENTS

	<b>Page</b>
1.INTRODUCTION.....	1
2. FINITE MIXTURE MODELS.....	7
1.1 MIXTURE MODELS IN PARAMETRIC CONTEXT .....	10
2.1.1 <i>Definition of the model</i> .....	10
2.1.2 <i>Clustering via mixture models</i> .....	11
2.2. FITTING MIXTURE MODELS VIA EM ALGORITHM.....	12
2.2.1 <i>General presentation of the EM algorithm</i> .....	13
2.2.2 <i>Formulation of the EM algorithm for mixture models</i> .....	15
2.2.3 <i>Information matrix using the EM algorithm</i> .....	17
2.2.4 <i>Starting values for EM Algorithm</i> .....	20
2.2.5 <i>Random Starting Values</i> .....	21
2.2.6 <i>Modified versions of the EM algorithm</i> .....	22
2.3 Choosing the number of clusters via model selection criteria.....	23
2.3.1 <i>Bayesian approaches for model selection</i> .....	24
2.3.2 <i>Strategy -oriented criteria</i> .....	25
2.3.3 <i>Partial Classification</i> .....	27
3. PCA – FMA – PGMM.....	29
3.1 <i>Principal Component Analysis</i> .....	29
3.2 <i>Single-Factor Analysis Model</i> .....	30
3.3 <i>EM Algorithm for a Single-Factor Analyzer</i> .....	31
3.4 <i>Mixtures of Factor Analyzers</i> .....	33
3.5 AECM ALGORITHM FOR FITTING MIXTURES OF FACTOR ANALYZERS.....	35
3.5.2 <i>AECM Framework</i> .....	35
3.5.2 <i>First Cycle</i> .....	36
3.5.3 <i>Second Cycle</i> .....	37
3.6 <i>Mixtures of Common Factor Analyzers (MCFA)</i> .....	39
3.7 <i>Parsimonious Gaussian Mixture Models</i> .....	41



	<b>Page</b>
4 APPLICATION.....	47
4.1 <i>Dimension Reduction</i> .....	47
4.2 <i>Choosing models according BIC criterion</i> .....	50
4.3 <i>Application of a two-component mixture factor analyzer with <math>q=6</math> factor</i> .....	51
4.4 <i>Application of a three-component mixture factor analyzer with <math>q=6</math> factor</i> .....	55
4.5 <i>Application of a two-component mixture factor analyzer with <math>q=8</math> factor</i> .....	58
5 Conclusion.....	62
6 References.....	64



## LIST OF TABLES

Table	Page
3.1 Parsimonious covariance structures derived from the mixture of factor analyzers model.....	43
4.1 Max-likelihood for models according to number of factors and number of groups .....	50
4.2 BIC for models according to number of factors and number of groups .....	51
4.3 Component probabilities, for two-component mixture factor analyzer with $q=6$ factors.....	52
4.4 Factor loading matrix, for two-component mixture factor analyzer with $q=6$ factors.....	52
4.5 Array of factor covariance matrix for two-component mixture factor analyzer with $q=6$ factors.....	53
4.6 Component probabilities for dataset with 500 variables, for a three-component mixture factor analyzer with $q=6$ factors.....	55
4.7 Factor loading matrix for dataset with 500 variables, for a three-component mixture factor analyzer with $q=6$ factors.....	56
4.8 Array of factor covariance matrix for dataset with 500 variables, for a three-component mixture factor analyzer with $q=6$ factors.....	56
4.9 Component probabilities, for dataset with 500 variables, for a two-component mixture factor analyzer with $q=8$ factors.....	59
4.10 Factor loading matrix for dataset with 500 variables, for dataset with 500 variables, for a two-component mixture factor analyzer with $q=8$ factors.....	59
4.11 Array of factor covariance matrix for dataset with 500 variables, for a two-component mixture factor analyzer with $q=8$ factors.....	60



## LIST OF FIGURES

<b>Figure</b>	<b>Page</b>
4.1 Plot of Log-likelihood for a two-component mixture factor analyzer with $q=6$ factor.....	51
4.2 Plot of error covariance matrix for clustering with two components and $q=6$ factors.....	53
4.3 Plot of clustering with two-components and $q=6$ factors.....	54
4.4 Plot of Log-likelihood for a three-component mixture factor analyzer with $q=6$ factor.....	55
4.5 Plot of error covariance matrix for clustering with three components and $q=6$ factors.....	57
4.6 Plot of clustering with three-components and $q=6$ factors.....	58
4.7 Plot of Log-likelihood for a two-component mixture factor analyzer with $q=8$ factor.....	59
4.8 Plot of error covariance matrix for clustering with two components and $q=8$ factors.....	60
4.9 Plot of clustering with two-components and $q=8$ factors.....	61









## 1. INTRODUCTION

As we move forward in time and as information technology improves, the fact is that data will be available for many more series and over an increasingly long span. While the availability of more data provides the advantage of understanding phenomena and anomalies better, we can also suffer from an information overload without some way to organize the data into an easy to interpret manner.

High-dimensional data are nowadays rule rather than exception in areas like information technology, bioinformatics or astronomy, as it is mentioned by Bühlmann and van de Geer (2011). The word “high-dimensional” refers to a situation where the number of unknown parameters which are to be estimated is one or several orders of magnitude larger than the number of samples in the data. We are unable to use classical statistical inference for high-dimensional problems. High-dimensional statistical inference is impossible without additional assumptions. A well-established framework for fitting many parameters is based on assuming structural smoothness, enabling estimation of smooth functions, as Bühlmann and van de Geer (2011) have mentioned.

As it is mentioned before, high-dimensional statistics refers to statistical inference when the number of unknown parameters  $p$  is several orders of magnitude larger than sample size  $n$ , that is:  $p \gg n$ . This encompasses supervised regression and classification models where the number of covariates is of much larger order than  $n$ , unsupervised settings such as clustering or graphical modeling with more variables than observations or multiple testing where the number of considered testing hypotheses is larger than sample size.

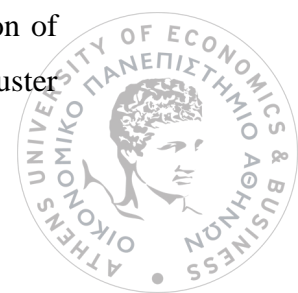
In an era with many high-throughput biological technologies, biomedical researchers are investigating more comprehensive aspects of cancer with ever-finer resolution. This not only results in large amount of data but also data with thousands of dimensions.



Li and Xu (2009), support that multivariate analysis is one of the most useful statistical tools in the analysis of biomedical data. It is about associating data matrices of  $n$  rows by  $p$  columns, with rows representing samples or patients or tissues and columns attributes, to certain response or outcome variables. Most of the times, the sample size  $n$  is much larger than the number of attributes  $p$ . Researchers discussed the theoretical properties of statistical models under the assumption of fixed  $p$  and infinite  $n$ . However, the advance of biological sciences and technologies has made the revolution to the process of investigations in cancer. The biomedical data collection has become much more automatic and much more extensive. We are in the era of  $p$  as a large fraction of  $n$ , or even much larger than  $n$ , which is a challenge for the classical statistical paradigm and calls for scalable solutions to the analysis of such high-dimensional data. In this volume, we will present the most known analytical approaches as well as systematic strategies to the analysis of correlated and high-dimensional data.

Bai and Ng (2007) mentioned that in recent years, theoretical and empirical researchers have focused on the analysis of large dimensional data. The early attention has primarily been on the use of factor models as a means of dimension reduction. But the volume of research, not only at the empirical but also at theoretical levels, has grown substantially. Empirical researchers support that it is useful to extract a few factors from a large number of series in many forecasting and policy exercises. Theoretical researchers tried to extend standard factor analysis to allow the size of both dimensions of a panel data set to increase. Now, we can understand better the theoretical implications of using estimated factors in both estimation and inference. Factor analysis plays a role not just in forecasting. Recently, the factor structure has been incorporated into regression analysis to deal with cross-sectionally correlated errors and endogeneity bias.

Bouveyron and Brunet (2014) claimed that clustering is a data analysis tool which aims to group data into several homogeneous groups. Researchers studied for years the clustering problem, which usually occurs in applications for which a partition of the data is necessary. Above all, more and more scientific domains require to cluster



data in order to understand or interpret the phenomenon that we want to study. Earliest approaches were based on heuristic or geometric procedures. They relied on dissimilarity measures between pairs of observations. A popular dissimilarity measure is based on the distance between groups, which was introduced by Ward for hierarchical clustering. In the same way, the k-means algorithm is also one of the most popular clustering algorithms among the geometric procedures. Clustering was also defined in a probabilistic framework, allowing to formalize the notion of clusters through their probability distribution. One of the most important advantages of this probabilistic approach is that the obtained partition can be interpreted from a statistical point of view.

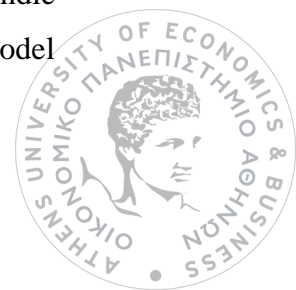
In our days, the measured observations in many scientific fields are frequently high-dimensional and clustering such data is a challenge for model-based methods. Indeed, model-based methods do not behave appropriate in high-dimensional spaces. They suffer from the well-known curse of dimensionality as Bellman (1957) claimed, and this is because model-based clustering methods are over-parameterized in high-dimensional spaces. What is more, in several applications, such as mass spectrometry or genomics, the number of available observations is small compared to the number of variables and such in this case the problem is more difficult. Since the dimension of observed data is usually higher than their intrinsic dimension, theoretically, it is possible to reduce the dimension of the original space without losing any information. This is why dimension reduction methods are usually used in practice in order to reduce the dimension of the data before the clustering step. Feature extraction methods, such as feature selection methods, or principal component analysis (PCA), are very popular. However, dimension reduction usually does not consider the classification task and provide a sub-optimal data representation for the clustering step. Indeed, dimension reduction methods imply an information loss which could have been discriminative.

To avoid the disadvantages of dimension reduction, researchers proposed several approaches to allow model-based methods to efficiently cluster high-dimensional data. This work proposes to review the alternatives to dimension reduction for dealing with high-dimensional data in the context of model-based clustering. Earliest



approaches include constrained and parsimonious models or regularization. More recently, subspace clustering techniques and variable selection techniques have been proposed as Bouveyron and Brunet (2014) claimed to overcome the limitations of previous approaches. Subspace clustering techniques are based on probabilistic versions of the factor analysis model. This modeling gives us the opportunity to cluster the data in low-dimensional subspaces without reducing the dimension. Conversely, variable selection techniques do reduce the dimension of the data but select the variables to retain regarding the clustering task. Both techniques turn out to be very efficient and we will discuss their practical use in this section.

This thesis contains four chapters. The opening chapter provides, as it is presented by McLachlan and Peel (2000), an overview of finite mixture models, as it was given much attention to the use of them as a device for clustering. Finite mixture models of distributions have provided a mathematical-based approach to the statistical modeling of a plethora of random phenomena. Over the years, they have continued to receive increasing attention, from both a practical and theoretical point of view, because of their usefulness as an extremely flexible method of modeling. Mixture models have been applied in fields such as engineering, astronomy, economics, biology, psychiatry, genetics, medicine, and marketing, among many other fields in the physical, biological, and society sciences. In these applications, finite mixture models underpin a plethora of technique. In major areas of statistics, including cluster and latent class analyses, image analysis, survival analysis and discriminant analysis, in addition to their more direct role in data analysis and inference of providing descriptive models for distributions. The usefulness of mixture distributions in the modeling of heterogeneity in a cluster analysis context is obvious. A finite mixture of normal densities with common variance (or covariance matrix in the multivariate case) can approximate arbitrarily any continuous distribution. Thus, mixture models provide a convenient semi-parametric framework in which to model unknown distributional shapes, whatever the objective, whether it be, say, density estimation or the flexible construction of Bayesian priors. A mixture model is able to model quite complex distributions through an appropriate choice of its components to represent accurately the local areas of support of the true distribution. It can thus handle situations where a single parametric family is not able to provide a satisfactory model



for local variations in the observed data. Inferences about the modeled phenomenon can be easily made from the mixture components, since the latter are chosen for their tractability.

Because of their flexibility, mixture models play a useful role in neural networks (Bishop, 1995, Section 5.9). With neural networks formed by the use of radial basis functions, the input data can be modeled by a mixture model. That is, the basic functions considered as the components of this mixture model after estimation by maximum likelihood from the input data. Then, the second-layer weights in the neural network can be estimated from the input data and their known outputs.

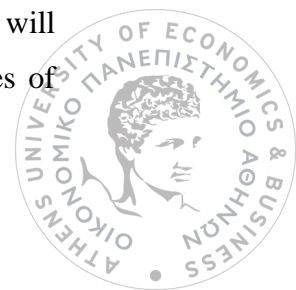
The third chapter has to do with factor analysis, which is commonly used for explaining data, in particular, correlations between variables in multivariate observations. Factor analysis is also used for dimensionality reduction, even if the method of principal component analysis is more widely used in this role. Although the effectiveness of these two methods can be widened by combining local models of them in the form of a finite mixture, as McLachlan and Peel (2000) mentioned.

We will also mention **Parsimonious Gaussian mixture models** as it is presented by McNicholas and Murphy (2008). Parsimonious Gaussian mixture models which are developed using latent Gaussian models are closely related to the factor analysis model. These models provide a modeling framework which introduce the mixture of probabilistic principal component analyzers and mixture of factor of analyzers models as special cases.

Initially, the basic underlying factor analysis and probabilistic principal components analysis models are described and it is reviewed the use of the EM algorithm to find maximum likelihood estimates for these models.

Then, we will introduce a class of eight parsimonious Gaussian mixture models which are based on the mixtures of factor analyzers model and the maximum likelihood estimates for the parameters in these models are found using an AECM algorithm.

In this chapter we will present mixtures of factor analyzers from the view of both a method for model-based density estimation for high-dimensional data (and hence for the clustering of such data) and a method for local dimensionality reduction. We will also try to examine the close link of mixtures of factor analyzers with mixtures of



probabilistic principal component analyzers. The mixtures of factor analyzers model enables a normal mixture model to be fitted to high dimensional data. The number of free parameters is controlled through the dimension  $q$  of the latent factor space. It allows an interpolation in model complexities from isotopic to full covariance structures with no restrictions.

Before the presentation of mixtures of factor analyzers, we will refer to the use of principal components for revealing group structure and dimension reduction

. In chapter four, a model-based approach to the clustering of microarray expression data received our attention, as it was approached by McLachlan et al. (2002), in particular, of tissue samples on a very large number of genes. The latter is a nonstandard problem in parametric cluster analysis because the dimension of the feature space (the number of genes) is typically much greater than the number of tissues. We will provide a feasible approach by first selecting a subset of the genes relevant for the clustering of the tissue samples by fitting mixtures of  $t$  distributions to rank the genes in order of increasing size of the likelihood ratio statistic for the test of two versus three components in the mixture model. The imposition of a threshold on the likelihood ratio statistic used in conjunction with a threshold on the size of a cluster allows the selection of a relevant set of genes. Although, even this reduced set of genes the most times will be too large for a normal mixture model to be fitted directly to the tissues, and this is why the use of mixtures of factor analyzers is exploited - to reduce effectively the dimension of the feature space of genes.

We will present the usefulness of the MCFA approach for the clustering of tissue samples on a data set on leukaemia tissues. For this data set, we will select relevant subsets of the genes that provide interesting clusterings of the tissues that are consistent with the external classification of the tissues.

In closing chapter, we will present which approach is mostly preferred according to bibliography and we will explain for which reasons we have chosen the number of factors  $q$  and the number of components  $g$  in chapter 4.



## Chapter 2

### Finite Mixture Model

Picard (2007) claimed that cluster analysis is primarily used to determine the inner structure of clustered data when there is no other information than the observed values. Clustering has drawn more attention over the years thanks to the emergence of new domains of application, such as astronomy, biology, physics and social sciences. As far as the clustering done in practice is concerned, most of it depends mainly on heuristic or distance-based procedures, such as hierarchical agglomerative clustering or iterative relocation procedures. Two major advantages are demonstrated in these methods: firstly, their construction is easily understood and secondly, the associated computational time is reasonable. However, their use seems to be limited due to the lack of their statistical basis, since typical questions in clustering, for instance, the number of clusters, can barely be theoretically handled by heuristic procedures.

Clustering methods which depend on probability models provide a foremost alternative to heuristic-based algorithms. The data in this context is considered as coming from a mixture of probability distributions, each representing a different cluster. Besides clustering purposes, finite mixtures of distributions have been applied to a great deal of statistical problems, such as discriminant analysis, image analysis and survival analysis. To this extent, finite mixture models have continued to focus more and more attention from both theoretical and practical points of view.

As it is mentioned by McLachlan and Peel (2000) we let  $Y_1, \dots, Y_n$  denote a random sample of size  $n$ , where  $Y_j$ , is a  $p$ -dimensional random vector with probability density function  $f(y_j)$  on  $R^p$ . In practice,  $Y_j$  contains the random variables corresponding to  $p$  measurements made on the  $j$ -th recording of some features on the phenomenon under study. We let  $Y = (Y_1^T, \dots, Y_n^T)^T$ , where the superscript  $T$  denotes vector transpose. Where possible, a realization of a random vector is denoted by the corresponding lower-case letter. For example,  $y = (y_1^T, \dots, y_n^T)^T$  denotes an observed random sample where  $y_j$ , is the observed value of the random vector  $Y_j$ . Although we are taking the feature vector  $Y_j$  to be a continuous random vector here, we can still view  $f(y_j)$  as a density in the case where  $Y_j$ , is discrete by the adoption



of counting measure. We suppose that the density  $f(y_j)$  of  $Y_j$  can be written in the form:

$$f(y_j) = \sum_{i=1}^g \pi_i f_i(y_j)$$

where  $f_i(y_j)$  is a component density of the mixture, and  $\pi_i$  the weight of population  $p$  (with the constraints  $0 < \pi_i < 1$  ( $p=1, \dots, g$ ) and  $\sum_{i=1}^g \pi_i = 1 = I$ ). In many applications the component densities are assumed to belong to some parametric family. In this case, they are specified as  $f(y_j; \theta_i)$ ; where  $\theta_i$  is the unknown vector of parameters of the postulated form for the  $i$ th component of the mixture. Let  $\psi = (\pi_1, \dots, \pi_{i-1}, \theta_1, \dots, \theta_i)$  denote the vector containing all the unknown parameters of the mixture.

Since we are interested in clustering it appears that one information is missing regarding the observed sample: the assignment of data points to the different clusters. A new random variable is introduced and noted  $Z_{jp}$ , where:

$$Z_{jp} = \begin{cases} 1, & \text{if data point } y_j \text{ belongs to population } p \\ 0, & \text{otherwise} \end{cases}$$

We suppose, as Picard (2007) does, that variables  $\{Z_1, \dots, Z_n\}$  are independent with ( $Z_j = \{Z_{j1}, \dots, Z_{jg}\}$ ) and that the conditional density of  $Y_j$  given  $\{Z_{ji} = 1\}$  is  $f(y_j; \theta_i)$ . Therefore variables  $Z_{ji}$  can be viewed as categorical variables that indicate the labeling of the data points. Thus  $Z_j$  is assumed to be distributed according to a multinomial distribution consisting of one draw on  $g$  categories with probabilities  $\pi_1, \dots, \pi_g$ :

$$\{Z_{j1}, \dots, Z_{jg}\} \sim M(1; \pi_1, \dots, \pi_g)$$

In terms of clustering, the  $i$ th mixing proportion can be viewed as the prior probability that one data point belongs to population  $p$ . The posterior probability of  $Z_{ji}$  given the observed value of  $y_j$  will be central for clustering purposes:



$$t_{ji} = Pr\{Z_{ji} = 1 | Y_j = y_j\} = \frac{\pi_i f(y_j; \theta_i)}{\sum_{m=1}^g \pi_m f(y_j; \theta_m)}$$

In order to formalize the incomplete data structure of mixture models, let  $X = (Y, Z)$  denote the complete data vector, whose only component being observed is  $Y$ . This reformulation clearly shows that mixture models can be viewed as a particular example of models with hidden structure such as hidden Markov models or models with censored data.

In order to formalize the incomplete data structure of mixture models, let  $X = (Y, Z)$  represent the complete data vector, whose only component being observed is  $Y$ . This reformulation proves definitely that mixture models can be considered as a special example of models with hidden structure such as hidden Markov models or models with censored data.

If the label of each data point was observed, the estimation of the mixture parameters would be straightforward since the parameters of each density component  $f(y_j; \theta_p)$  could be estimated only via the data points from population  $p$ . However, the categorical variables are hidden, and the estimation can depend solely on the observed data  $Y$ . The main reason for the considerable work on estimation methodology for mixtures is that explicit formulas for parameter estimates are not available in a closed form and as a result it leads to the need for iterative estimation procedures.

A wide variety of techniques can be used to handle fitting mixture distributions. According to Picard (2007), such graphical methods are the following: the method of moments, maximum likelihood and Bayesian approaches. Important progress in the fitting of mixture models, especially via the maximum likelihood method, has been made quite recently, specifically it has been since 40 years. The publication of *Dempster et al. (1977)* and the *Introduction of the EM algorithm* played a key role in that kind of progress.

The iterative computation of maximum likelihood estimators is the purpose of the EM algorithm when observations can be viewed as incomplete data. The main idea of the EM algorithm is to associate a complete data model with the incomplete structure that is observed so as to make the computation of maximum likelihood estimates less complex. Likewise, a complete data likelihood is associated with the complete data



model. The EM algorithm exploits the simpler MLE computation of the complete data likelihood to optimize the observed data likelihood. The EM algorithm and its general properties will be described later. In spite of a great variety of successful applications and the significant work on its properties, the EM algorithm displays two intrinsic limitations: it appears to be slow to converge and as many iterative procedures, it is prone to the initialization step. That's why modified versions of the EM algorithm have been developed.

Once the mixture model has been specified and its parameters have been estimated, one basic question remains: "How many clusters?". Mixture models present a main advantage compared with heuristic cluster algorithms in which there is no established method to define the number of clusters. With the underlying probability model, the problem of choosing the number of components can be reformulated as a statistical model choice problem. Testing for the number of components in a mixture appears to be not an easy job since the typical likelihood ratio test does not hold for mixtures. In contrast, criteria based on penalized likelihood, for instance the Bayesian Information Criterion (BIC) have been successfully applied to mixture models.

Nonetheless, it seems that those criteria do not take into account the specific objective of mixture models in the clustering context. This accounts for the construction of classification-based criteria.

## 2.1 MIXTURE MODELS IN PARAMETRIC CONTEXT

### 2.1.1 Definition of the model

We let  $Y = \{Y_1, \dots, Y_n\}$  a random sample of size  $n$ , where  $Y_j$  is a vector of  $R^p$  with probability density function  $f(y_j)$ . In the mixture model context the density of  $Y_j$  is supposed to be a mixture of  $P$  parametric densities such that:

$$f(y_j; \psi) = \sum_{i=1}^g \pi_i f(y_j; \theta_i) \quad (2.1)$$

with the constraint  $\sum_{i=1}^g \pi_i = 1$ ,  $P$  being fixed. Coefficients  $\pi_i$  can be viewed as the weights of the  $i$ th component of the mixture, which is characterized by parameter  $\theta_i$ .

$$\psi = (\pi_1, \dots, \pi_{i-1}, \theta_1, \dots, \theta_i)$$



represents the vector of parameters of the model.

Mixture models are reformulated as an incomplete data problem since the assignment of the observed data is unknown. If we note  $X_j = \{Y_j, Z_j\}$ , the complete data vector whose only component being observed is  $Y_j$ , its density function is then:

$$g(x_j; \psi) = \prod_{i=1}^g [\pi_i f(y_j; \theta_i)]^{z_{ji}} \quad (2.2)$$

### 2.1.2 Clustering via mixture models

Questions about clustering may arise after the mixture model has already been fitted. In the first case, the reason for fitting a mixture model was to obtain a reliable model for the distribution of data. If this were accomplished by the fitting of, say, a three-component mixture model, we would want to see if the three components can be identified with three externally existing groups.

The clustering of the data at hand is the main aim of the analysis. The mixture model is being used exclusively as a device for representing any grouping that might be related to the data.

The usage of mixture models in the clustering context aims at a partition of the data into  $g$  groups, with  $g$  being fixed. The populations' weights stand for prior probabilities of belonging to a given population.  $Pr\{Z_{ji} = 1\} = \pi_i$  represents the probability to classify one data point to population when the only information given about the data is the weights of each group.

In the complete data specification the clustering procedure aims at recovering the associated label variables  $z_1, \dots, z_n$  having observed  $y_1, \dots, y_n$ . After the mixture model has been fitted and its parameter has been estimated, a probabilistic clustering of the observations is provided in terms of their posterior probabilities of component membership:



$$\widehat{\tau}_{ji} = Pr\{Z_{ji} = 1 | Y_j = y_j\} = \frac{\widehat{\pi}_i f(y_j; \widehat{\theta}_i)}{\sum_{m=1}^g \widehat{\pi}_m f(y_j; \widehat{\theta}_m)}$$

Probabilities  $\widehat{\tau}_{j1}, \dots, \widehat{\tau}_{jg}$  are the estimated probabilities that data point  $y_j$  belongs to the first, second, ...  $g^{th}$  component of the mixture. Instead of misclassification results, each data point can be assigned to a particular population with the maximum a posteriori rule (MAP)

$$\widehat{z}_{ji} = \begin{cases} 1, & \text{if } i = \operatorname{argmax}\{\widehat{\tau}_{jm}\} \\ 0, & \text{otherwise} \end{cases}$$

## 2.2. FITTING MIXTURE MODELS VIA EM ALGORITHM

According to Picard (2007), even with the advent of high-speed state-of-the-art computers, researchers of the past were somehow hesitant to fit mixture models to data of more than one dimension, probably due to the fact that they did not fully understand the issues that arise with their fitting. The presence of multiple maxima is included in the mixture likelihood function and the unboundedness of the likelihood function is included in the case of normal components with unequal covariance matrices. However, the difficulties about these computational issues were properly understood and successfully dealt with in the course of time and as a result there has been an increase in using mixture models in practice.

The estimation problem of the parameters of a mixture can be dealt with by a variety of methods from graphical to Bayesian methods. However, the maximum likelihood method seems to be the one which stands out of the rest, mostly because of the existence of a related statistical theory. Given a sample of  $n$  independent observations from a mixture defined in (2.1), the likelihood function is:

$$L(y; \psi) = \prod_{j=1}^n \left\{ \sum_{i=1}^g \pi_i f(y_j; \theta_i) \right\}$$



The particularity of mixture models is that the maximization of the likelihood which is described above with respect to  $\psi$  is not straightforward and requires iterative procedures. EM algorithm arose a great interest in the use of finite mixture distributions to model heterogeneous data. This is due to the fact that the fitting of mixture models by maximum likelihood is a typical instance of a problem that is considerably made less complex by the EM's conceptual unification of maximum likelihood (ML) estimation from data that can be considered as being not fully developed.

### 2.2.1 General presentation of the EM algorithm

In the incomplete data formulation of mixture models Picard (2007) noted  $X$  the complete data sample space from which  $x$  arises,  $Y$  the observed sample space and  $Z$  the hidden sample space. It follows that  $X = Y \times Z$  and  $x = (y; z)$ . The density of the observed data  $X$  can be written in the form:

$$g(x; \psi) = f(y; \psi)k(z|y; \psi),$$

where  $f(y; \psi)$  is the density of the observed data and  $k(z|y; \psi)$  is the conditional density of the missing observations given the data. As a result we have the definition of different likelihoods: the observed/incomplete-data likelihood  $L(y; \psi)$  and the unobserved/complete-data likelihood  $L^c(x; \psi)$ . These likelihoods are linked with the relationship:

$$\log L^c(x; \psi) = \log L(y; \psi) + \log k(z|y; \psi)$$

with

$$\log L^c(x; \psi) = \sum_{j=1}^n \log g(x_j; \psi)$$

and



$$\log k(z|y; \psi) = \sum_{j=1}^n \sum_{i=1}^g z_{ji} \log E\{Z_{ji}|Y_j = y_j\}$$

Since the hidden variables are not observed, the EM machinery includes the indirect optimization of the incomplete-data likelihood through the iterative optimization of the conditional expectation of the complete-data likelihood using the current fit for  $\psi$ . If we note  $\psi^{(h)}$  the value of the parameter at iteration  $h$ , it follows that:

$$\log L(y; \psi) = Q(\psi; \psi^{(h)}) - H(\psi; \psi^{(h)}) \quad (2.3)$$

with conventions:

$$Q(\psi; \psi^{(h)}) = E_{\psi^{(h)}}\{\log L^C(X; \psi)|Y\}$$

$$H(\psi; \psi^{(h)}) = E_{\psi^{(h)}}\{\log k(Z|Y; \psi)|Y\}$$

where  $E_{\psi^{(h)}}\{\cdot\}$  denotes the expectation operator, taking the current fit  $\psi^{(h)}$  for  $\psi$ .

The EM algorithm includes two steps:

- E-step: calculate  $Q(\psi; \psi^{(h)})$
- M-step: choose  $\psi^{(h+1)} = \operatorname{argmax} \{Q(\psi; \psi^{(h)})\}$ .

The E- and M- steps are repeated alternatively until the difference:  $|\psi^{(h+1)} - \psi^{(h)}|$  changes by an arbitrarily small amount. A different way of stopping rule could be the difference of log-likelihoods between two steps  $|\log L(y; \psi^{(h+1)}) - \log L(y; \psi^{(h)})|$ . Although if the log-likelihood is "flat" with respect to this difference can be stable whereas parameter  $\psi^{(h)}$  keeps changing.

The key property of the EM algorithm established by Dempster et al. (1977) is that the incomplete data log-likelihood increases after each iteration of the algorithm. The proof of this theorem is based on the definition of the M-step that ensures

$$Q(\psi; \psi^{(h+1)}) \geq Q(\psi; \psi^{(h)})$$



while the application of the Jensen inequality gives

$$H(\psi; \psi^{(h+1)}) \leq H(\psi; \psi^{(h)})$$

Put together and considering relation 2.3, these inequalities ensure the monotonicity of the likelihood sequence:

$$\log L(y; \psi^{(h+1)}) \geq \log L(y; \psi^{(h)})$$

This inequality proves that the EM sequence of likelihoods must converge if the likelihood is bounded above.

### 2.2.2 Formulation of the EM algorithm for mixture models

When applied to the special case of mixture models the log-likelihoods are written in the form:

$$\log L(y; \psi) = \sum_{j=1}^n \log f(y_j; \psi) = \sum_{j=1}^n \log \left\{ \sum_{i=1}^g \pi_i f(y_j; \theta_i) \right\}$$

$$\log L^C(x; \psi) = \sum_{j=1}^n \log g(x_j; \psi) = \sum_{j=1}^n \sum_{i=1}^g z_{ji} \log \{ \pi_i f(y_j; \theta_i) \}$$

Since the complete data log-likelihood is linear in the unobservable data  $z_{ji}$ , the E step is associated with the computation of the conditional expectation of the missing information given the observed data  $y_j$ , using the current fit  $\psi^{(h)}$  for  $\psi$ . It gives

$$Q(\psi; \psi^{(h)}) = \sum_{j=1}^n \sum_{i=1}^g E_{\psi^{(h)}} \{ Z_{ji} | Y_j = y_j \} \log \{ \pi_i f(y_j; \theta_i) \}$$

with



$$E_{\tau_{\psi^{(h)}}} = \{Z_{ji}|Y_j = y_j\} = Pr\{Z_{ji} = 1|Y_j = y_j\} = \tau_{ji}^{(h)}$$

and

$$\tau_{ji}^{(h)} = \frac{\pi_i^{(h-1)} f(y_j; \theta_i^{(h-1)})}{\sum_{m=1}^g \pi_m^{(h-1)} f(y_j; \theta_m^{(h-1)})}$$

$$Q(\psi; \psi^{(h)}) = \sum_{j=1}^n \sum_{i=1}^g \tau_{ji}^{(h)} \log\{\pi_i f(y_j; \theta_i)\}$$

As for the M-step, it is associated with the global maximization of  $Q(\psi; \psi^{(h)})$  with respect to  $\psi$  to generate an updated estimate  $Q(\psi; \psi^{(h+1)})$ .

In terms of finite mixture models, the estimation of the mixing proportions is achieved through constrained maximization of the incomplete-data log-likelihood which gives:

$$\hat{\pi}_i^{(h+1)} = \frac{\sum_{j=1}^n \tau_{ji}^{(h)}}{n}$$

This estimator has a natural interpretation: it summarizes the contribution of each data point  $y_j$  to the  $i$ th component of the mixture through its posterior probability of membership. As far as the updating of  $\theta$  is concerned, it is obtained as an appropriate root of the weighted likelihood equations

$$\sum_{j=1}^n \sum_{i=1}^g \tau_{ji}^{(h)} \frac{\partial \log f(y_j; \theta_i)}{\partial \theta} = 0 \quad (2.4)$$



### 2.2.3 Information matrix using the EM algorithm

Geoffrey McLachlan and David Peel (2000) point out that one initial criticism of the EM algorithm was that it does not automatically provide an estimate of the covariance matrix of the MLE, as other procedures such as Newton-type methods do. A number of methods, which depends on the observed information matrix  $I(\hat{\Psi}; y)$ , has already been suggested.

Since the parameters of the mixture have been estimated through maximum likelihood, a natural question is to assess the standard errors of the estimator  $\hat{\psi}$ . This can be done by evaluating the expected information matrix

$$I(\psi) = E_Y \left\{ \frac{-\partial^2}{\partial \psi \partial \psi^T} \log L(Y; \psi) \right\},$$

with  $\log L(Y; \psi)$  being the incomplete-data likelihood estimated on the available observations, and  $E_Y\{\cdot\}$  representing the expectation operator regarding the random variable  $Y$ .

From a practical point of view, this quantity is mostly estimated by the observed information matrix calculated at  $\hat{\psi}$ ,  $I(\hat{\psi}, y)$  with the relationship

$$I(\psi) = E_Y\{I(\psi; Y)\}$$

According to Efron and Hinkley (1978),  $I(\psi; Y)$  will be denoted as the incomplete-data observed information matrix since the data  $Y$  is considered as incomplete within the EM framework.

The use of the EM algorithm is usually motivated by the analytic form of the observed-data likelihood, whose gradient or curvature matrices are difficult to derive analytically. Since there has been a solution thanks to the missing-data framework of EM, the derivation of the information matrix  $I(\psi; y)$  can be made less complex using the missing information principle.



### **Missing information principle**

If the formulation of mixtures is viewed as a missing-data problem, we can describe the complete-data observed information matrix based on the complete data log-likelihood:

$$I^c(\psi; \chi) = \frac{-\partial^2}{\partial \psi \partial \psi^T} \log L^c(\chi; \psi)$$

since the incomplete data and the complete data likelihood are linked by definition:

$$\log L(y; \psi) = \log L^c(x; \psi) - \log k(z|y; \psi)$$

on differentiating both sides twice with respect to  $\psi$ , we have

$$I(\psi; y) = I^c(\psi; \chi) - I^m(\psi, z)$$

where

$$I^m(\psi, z) = \frac{-\partial^2}{\partial \psi \partial \psi^T} \log k(z|y; \psi)$$

is the missing-data observed information matrix. We can consider this term as the "missing information", since only  $y$  and not  $z$  has been observed.

The complete-data is not fully observed, so we take the conditional expectation of both sides over  $Y$  that yields to:

$$I(\psi; y) = E_{z|y} \{I^c(\psi; X)\} - E_{z|y} \{I^m(\psi, Z)\} \quad (2.5)$$

Then the problem is to formulate the conditional expectations of  $I^c(\psi; \chi)$  and  $I^m(\psi, z)$  in directly computable terms within the EM framework.



## Extracting the observed information matrix in terms of the complete-data

### Likelihood

The likelihood function for  $\Psi$ , which comes from the observed data  $y$ , is denoted by  $L(\Psi)$ , while  $L^c(\Psi)$  represents the complete data likelihood function for  $\Psi$  that could be formed from the complete vector  $y_c$  if it were utterly observable.

The (incomplete-data) score statistic is as shown below:

$$S(y; \psi) = \frac{\partial}{\partial \psi} \log L(y; \psi)$$

while the complete-data score statistic is given:

$$S^c(x; \psi) = \frac{\partial}{\partial \psi} \log L^c(x; \psi)$$

Louis (1982) gives a formulation of the missing information matrix, in the form:

$$E_{Z|Y}\{I^m(\psi, Z)\} = E_{X|Y}\{S^c(X; \psi)S^c(X; \psi)^T\} - S(y; \psi)S(y; \psi)^T$$

meaning that the all the conditional expectations calculated in 2.4 can be computed in the EM algorithm only using the conditional expectation of the gradient and curvature of the complete-data likelihood.

Since  $S(y; \psi) = 0$  for  $\psi = \hat{\psi}$  Formula 2.4 is restated as:

$$I(\hat{\psi}; y) = E_{X|Y}\{I^c(\psi; X)\}_{\psi=\hat{\psi}} - E_{X|Y}\{S^c(X; \psi)S^c(X; \psi)^T\}_{\psi=\hat{\psi}}.$$

Hence the observed information matrix of the initial incomplete-data problem can be computed as the conditional moments of the gradient and curvature matrix of the complete-data likelihood introduced in the EM framework.



### 2.2.4 Starting values for EM Algorithm

As it was previously mentioned EM algorithm is started from initial value of  $\Psi$ ,  $\Psi^{(0)}$ . Therefore, to put it into practice, we have to define a value for  $\Psi^{(0)}$ . Seidel, Mosler, and Alker (2000) have proved how different starting strategies and stopping rules can result in quite different estimates regarding fitting exponential components through the EM algorithm.

The situation is slow because of the convergence with the EM algorithm. The situation could worsen by a poor choice of  $\Psi^{(0)}$ . In some cases where the likelihood is unbounded on the edge of the parameter space, the sequence of estimates  $\{\psi^{(h)}\}$  which comes from the EM algorithm might diverge if  $\Psi^{(0)}$  is chosen too close to the boundary. One more problem regarding mixture models is that the likelihood equation will usually have multiple roots corresponding to local maxima and as a result the EM algorithm should be applied from a wide range choice of starting value in any search for all local maxima.

In the absence of the observed value of any known consistent estimator of  $\Psi$  or any other information, an apparent choice for the root of the likelihood equation is the one corresponding to the largest of the local maxima located. Nevertheless, it does not mean that this choice defines the sequence of roots of the likelihood equation that is consistent and asymptotically efficient.

For independent data in the case of mixture models, the effect of the E-step is to update the posterior probabilities of component membership. Hence an alternative approach is to perform the first E-step by specifying a value  $\tau_j^{(0)}$  for  $\tau(y_j; \Psi)$  for each  $j(j=1, \dots, n)$ , where

$$\tau(y_j; \Psi) = (\tau_1(y_j; \Psi), \dots, \tau_g(y_j; \Psi))^T$$

Is the vector which includes the  $g$  posterior probabilities of the component membership for  $y_j$ . The latter is usually undertaken by setting  $\tau_j^{(0)} = z_j^{(0)}$  for  $j=1, \dots, n$ , where

$z^{(0)} = (z_1^{(0)T}, \dots, z_n^{(0)T})^T$  represents an initial partition of the data into  $g$  groups.

For higher dimensional data, an initial value  $z^{(0)}$  for  $z$  may be obtained through the use of a clustering algorithm, such as k-means or, say, an hierarchical procedure if  $n$  is not too large.



### 2.2.5 Random Starting Values

McLachlan and Peel (2000) demonstrate another way of specifying an initial partition  $z^{(0)}$  of the data according to which the data is randomly divided into  $g$  groups corresponding to the  $g$  components of the mixture model. In other words, an integer between 1 and  $g$ , both inclusive, is randomly generated for each observation  $y_j$ . If this random integer equals  $h$ , then we set the  $i$ th element of  $z_j^{(0)}$  equal to one for  $i = h$  and equal to zero for  $i \neq h$  ( $i=1, \dots, g$ )

Most times, the EM algorithm would be applied from a number of random starts. Random starts affects the central limit theorem as it tends to have similar component parameters in large samples. If we first select a small random subsample from the data, which is then randomly assigned to the  $g$  components, then we can reduce this effect. The first M-step is then made on the basis of the subsample. The subsample has to be sufficiently large to guarantee that the first M-step can produce a nondegenerate estimate of the parameter vector  $\Psi$ .

Another way of specifying a random start, at least in the context of  $g$  normal components with means  $\mu_i$  and covariance matrices  $\Sigma_i$ , is to generate the  $\mu_i^{(0)}$  independently in random as in the example below:

$$\mu_1^{(0)}, \dots, \mu_g^{(0)} \sim N(\bar{y}, V)$$

Where  $\bar{y}$  is the sample mean and

$$V = \sum_{j=1}^n (y_j - \bar{y})(y_j - \bar{y})^T / n$$

is the sample covariance matrix of the observed data. With this method, there is more variation between the initial values  $\mu_i^{(0)}$  for the component means  $\mu_i$ , than with a random partition of the data into  $g$  groups, and it is also computationally less demanding.

The component-covariance matrices  $\Sigma_i$  and the mixing proportions  $\pi_i$  can be specified as

$$\Sigma_i^{(0)} = V \text{ and } \pi_i^{(0)} = \frac{1}{g} \quad (p = 1, \dots, g).$$



As illustrated in McLachlan and Basford (1988), a key factor in the fitting of a mixture model is the accuracy of the estimate of the vector of the mixing proportions.

### 2.2.6 Modified versions of the EM algorithm

Although the EM algorithm has appealing features, it also has some well documented shortcomings: the resulting estimate  $\hat{\psi}$  can strongly be determined by the starting position  $\psi^{(0)}$ , the rate of convergence can be slow and it can provide a saddle point of the likelihood function rather than a local maximum. For the reasons mentioned above, a great deal of authors have suggested modified versions of the EM algorithm: deterministic improvements (Louis (1982), Meilijson (1989), Green (1990)), and stochastic modifications (Broniatowski et al. (1983) Celeux and Diebolt (1985) Wei and Tanner (1990), Delyon et al. (1999)).

Broniatowski et al. (1983) proposed a Stochastic EM algorithm (SEM) which stands for an appealing alternative to EM. The motivation of the simulation step (S-step) relies on the Stochastic Imputation Principle, where the purpose of the S-step is to fill-in for the missing data  $z$  with a single draw from  $(z|y; \psi^{(h)})$ . This imputation of  $z$  is based on all the information at hand about  $\psi$  and provides a pseudo complete sample. Specifically, the current posterior probabilities  $\tau_{ji}^{(h)}$  are used in the S-step wherein a single draw from distribution  $M_i(1; \tau_{j1}^{(h)}, \dots, \tau_{jg}^{(h)})$  is used to assign each observation to one of the components of the mixture. The deterministic M-Step and the stochastic S-Step create a Markov Chain  $\psi^{(h)}$ , which converges to a stationary distribution under mild conditions. Practically, several iterations are needed as a burn in period to allow  $\psi^{(h)}$  to approach its stationary regime. In mixture models 100-200 iterations are usually used for burn in.

This stochastic step can be thought of as a random perturbation of the sequence  $\psi^{(h)}$  generated by EM. This perturbation prevents the algorithm from staying near an changeable fixed point of EM, and prevents stable fixed points corresponding to insignificant local maxima of the likelihood. The Stochastic EM algorithm offers an intriguing alternative to the limitations of EM regarding local maxima and starting values.



Moreover, there are some other stochastic versions of the EM algorithm which have been proposed as well. For instance, the Stochastic Annealing EM algorithm (SAEM, Celeux and Diebolt (1992)), which is a modification of SEM, the Monte Carlo EM (Wei and Tanner (1990)), which replaces analytic computation of the conditional expectation of the complete-data log-likelihood by a Monte Carlo approximation, and a stochastic approximation of EM (Delyon et al. (1999)). However, empirical studies from Dias and Wedel (2004) and Biernacki et al. (2003) support the practical use of SEM in the context of mixture models because of the simplicity of implementation compared with Monte Carlo-based improvements, of its quick rate of convergence, and of its property to avoid deceitful local maximizers.

### 2.3 Choosing the number of clusters via model selection criteria

According to Picard (2007), the first thing that is usually asked by/to the analyst is to choose the number of clusters. Two approaches can be taken into account in order to answer the question above. According to the first approach, this number can be fixed and different classifications can be proposed. Every clustering method (heuristically or model-based) can be run for a fixed number of groups, so this strategy can be applied to any method. Nonetheless, the question can be to score different classifications with different numbers of clusters. In the model-based context, the choice of the number of clusters can be developed as a model selection problem, and it can be performed with a penalized criterion, such as:

$$\log L_g(y; \hat{\psi}) - \beta \cdot P_{penalized}$$

With  $L_g(y; \hat{\psi})$  being the observed data log-likelihood for a mixture with P clusters, calculated at  $\psi = \hat{\psi}$ ,  $\beta$  a positive constant and  $P_{penalized}$  an increasing function with respect to the number of clusters.



### 2.3.1 Bayesian approaches for model selection

As described above, in the context of segmentation methods, the model selection aims at selecting a candidate model  $m_i$  among a finite collection of models  $\{m_1, \dots, m_l\}$ , in order to estimate function  $f$  from which the data  $Y = \{Y_1, \dots, Y_n\}$  is generated. Each model is marked by a density  $g_{m_i}$  whose parameters  $\psi_i$  belong to the dimension  $v_i$ .

As for the Bayesian context,  $\psi_i$  and  $m_i$  are viewed as random variables with prior distributions noted  $Pr\{m_i\}$  and  $Pr\{\psi_i|m_i\}$  for  $\psi_i$  when model  $m_i$  is fixed. This formulation is flexible since extra information can be modeled through *prior* distributions, and if there is not any information available, then a non-informative prior can be used. The Bayesian Information Criterion (BIC) developed by Schwartz (1978) aims at selecting the model which maximizes the posterior probability  $Pr\{m_i|Y\}$ . Using the Bayes formula:

$$Pr\{m_i|Y\} = \frac{Pr\{Y|m_i\}Pr\{m_i\}}{Pr\{Y\}}$$

and considering the case where the prior distribution  $Pr\{m_i\}$  is non informative, the search for the best model only needs the computation of distribution  $Pr\{Y|m_i\}$ , which is the integrated likelihood of the data for model  $m_i$ . This distribution can be approximated using the Laplace approximation method (see Lebarbier and Mary-Huard (2004) for more details), which yields to the following penalized criterion:

$$BIC_i = -2Pr\{Y|m_i\} \approx -2 \log g_{m_i}(Y; \widehat{\psi}_i) + v_i \times \log(n)$$

Where  $\widehat{\psi}_i$  is the maximum likelihood estimator of  $\psi_i$ . The BIC is used to assess a score to each model  $m_i$  and the selected model is such that:

$$\widehat{m}_{BIC} = Argmax BIC_i$$

In an interesting way, regularity conditions for BIC do not apply to mixture models, because the estimates of some mixing proportions can be on the boundary of the parameter space. Despite this, there is important practical support for its use in this context (see Fraley and Raftery (1998) for instance). Some other approaches have



been considered for the Bayesian model selection as well (see Kass and Raftery (1995) for a complete review on Bayes Factors for instance). However, the BIC has drawn much attention, both for its simplicity of implementation and for its statistical properties. Gassiat and Dacunha-Castelle (1997) have demonstrated that the use of BIC results in a consistent estimator of the number of clusters.

### 2.3.2 Strategy -oriented criteria

There are also some other criteria, which have been defined for the special case of mixture models. Those criteria can be based on Bayesian methods, on the entropy function of the mixture, or on information theory. McLachlan and Peel (2000) made a complete review on the construction of those criteria. In order to determine the "best" criterion, there has been an extensive use of Empirical comparisons of those criteria. Biernacki et al. (2000) noted that the use of the BIC can lead to an overestimation of the number of clusters no matter the clusters separation. Furthermore, estimating the "true" numbers of clusters, which is the objective of the BIC, is not inevitably necessary in a practical context. For the reasons mentioned above, Biernacki et al. (2000) suggest a new criterion, the Integrated Classification Criterion (ICL) which takes into account the clustering objective of mixture models. The main steps of the construction of ICL are presented in the paragraph below:

In a mixture model context, the integrated likelihood is noted  $f(y|m_p)$  for a model  $m$  with  $P$  clusters. It is calculated such that:

$$f(y|m_p) = \int_{\Psi_p} f(y|m_p, \psi)h(\psi|m_p)d\psi$$

with

$$f(y|m_p, \psi) = \prod_{t=1}^n f(y_t|m_p, \psi),$$

$\Psi_p$  being the parameter space of model  $m_p$ , and  $h(\psi|m_p)$  a non-informative prior distribution on  $\Psi$ . According to the authors, instead of considering the incomplete-data integrated likelihood for which the BIC approximation is not valid, it is



preferable to use the complete-data integrated likelihood or integrated classification likelihood:

$$f(y, z|m_p) = \int_{\Psi_p} f(y, z|m_p, \psi)h(\psi|m_p)d\psi$$

with

$$f(y, z|m_p, \psi) = \prod_{t=1}^n \prod_{p=1}^P \{\pi_p f(y_t; \theta_p)\}^{z_{tp}}$$

Then the idea is to isolate the contribution of the missing data  $z$  by conditioning on  $z$ , and it follows that:

$$f(y, z|m_p) = f(y|z, m_p) \int(z|m_p),$$

Provided that  $h(\psi, m_p) = h(\theta|m_p)h(\pi|m_p)$ .

The authors emphasize that the BIC approximation is valid for the term  $f(y, z|m_p)$ , such that:

$$\log f(y, z|m_p) \cong \max \log f(y, z|m_p, \theta) - \frac{\lambda_p}{2} \log(n),$$

Where  $\lambda_p$  is the number of free components in  $\theta$ . Note that the parameter  $\theta$  which maximizes  $\log f(y, z|m_p, \theta)$  is not the maximum likelihood estimator. Nevertheless, the authors propose to use the maximum likelihood estimator as an approximation.

As for term  $f(z|m_p)$  it can be directly calculated using a Dirichlet prior  $D(\delta, \dots, \delta)$  on proportion parameters. It follows that:

$$f(z|m_p) = \int \pi_1^{n_1}, \dots, \pi_p^{n_p} \frac{\Gamma(P\delta)}{\Gamma(\delta)^P} \sum_p \pi_p = 1 d\pi$$



With  $n_p$  being the number of data points belonging to cluster  $p$ , the parameter  $\delta$  is fixed at  $1/2$  which corresponds to the Jeffreys non-informative distribution for proportion parameters.

The last steps of the construction of ICL includes the replacing the missing data  $z$  which is unknown by the recovered label variables  $\tilde{z}$  using a MAP rule. Then an approximation of  $f(z|m_p)$  is given when  $n$  is large. It follows that:

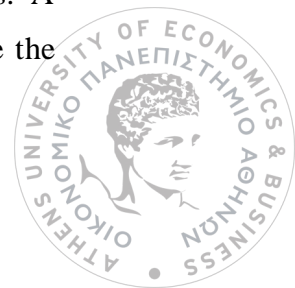
$$ICL_{(m_p)} = \max \log f(y, \tilde{z}|m_p, \psi) - \frac{v_p}{2} \log(n)$$

with  $v_p$  the number of free parameters for model  $m_p$ . Hence, the ICL criterion is an "la BIC" approximation of the completed log-likelihood or classification log likelihood. This criterion has been shown to lead to a more sensible division of the data, compared with BIC, because it considers the classification results to score each model. Picard (2007) alleged that the performance of ICL has been examined through real and simulated data sets. ICL tends to select fewer clusters which offer good clustering results in real situations, compared with BIC, which tends to select extremely too many clusters. When the data is simulated, ICL tends to select fewer clusters if the groups are not well separated. On the contrary, BIC finds out the real number of classes. Theoretically, there has been no result yet for the properties of ICL.

### 2.3.3 Partial Classification

McLachlan and Peel (2000) argued that in situations where the observed data contains some observations whose component of origin is known, the observed data  $y_1, \dots, y_n$  contain some data that is classified in accordance with the components of the mixture model.  $y_j$  ( $j = 1, \dots, m$ ) denote the  $m$  ( $m < n$ ) classified observations; that is, for these  $y_t$  the associated component indicator vectors  $z_j$  are known.

This may arise in situations where the components correspond to externally existing groups, and some of the observed data has been classified in terms of these groups. A discriminant rule has been formed from the classified data and the aim is to use the



subsequent unclassified data to enhance the performance of the rule by forming it on the basis of the combined classified documents which are already available.

Taking into consideration both the classified and unclassified data, the estimation can be undertaken in a straightforward manner by maximum likelihood via the EM algorithm. The equation

$$\sum_{i=1}^g \sum_{j=1}^n \tau_i(y_j; \Psi^{(h)}) \partial \log \frac{f_i(y_j; \theta_i)}{\partial \xi} = 0$$

for the update  $\xi^{(h+1)}$ , containing the parameters in the component densities of the mixture model still applies in the presence of some classified data, except that we use the known value of the component indicator variable  $z_{ij}$  instead of its currently evaluated expectation  $\tau_i(y_j; \psi^{(h)})$  ( $i = 1, \dots, g$ ). For the update of the  $i$ th mixing proportion  $\pi_i$ , assuming that the classified data has been obtained by sampling from the mixture. If the classified data provides no information on the  $\pi_p$ , then the updated estimate of  $\pi_i$  is given by:

$$\pi_i^{(h+1)} = \sum_{j=m+l}^n \frac{\tau_i(y_j; \psi^{(h)})}{n - m}$$

The presence of data of known origin with respect to each component of the mixture facilitates Maximum likelihood estimation. There may be singularities in the likelihood on the edge of the parameter space if there are any component densities that are multivariate normal with unequal covariance matrices. Despite this, there will be no singularities if there are more than  $p$  classified observations available from each component. Based exclusively on the classified data, the MLE of  $\Psi$  is an obvious choice of a starting point in the presence of classified data.



## Chapter 3

### PCA – FMA – PGMM

Factor analysis, which is mostly, used for explaining data, in particular, correlations between variables in multivariate observations. Factor analysis is also used for dimensionality reduction, even if the method of principal component analysis is more widely used in this role. Although the effectiveness of these two methods can be widened by combining local models of them in the form of a finite mixture, as McLachlan and Peel (2000) mentioned.

#### 3.1 *Principal Component Analysis*

As it is mentioned by Geoffrey McLachlan and David Peel (2000), in exploring high-dimensional data sets for group structure, we typically rely on “second-order” multivariate techniques, in particular, principal component analysis (PCA) and the upcoming discussions for the excellent account of available exploratory multivariate techniques. Here we will discuss a PCA on the sample covariance matrix

$$V = \sum_{j=1}^n (y_j - \bar{y})(y_j - \bar{y})^T / n$$

We let  $a_1, \dots, a_p$  be the unit eigenvectors, corresponding to the eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  of  $V$ . If the variables are measured on different scales, we may use the sample correlation matrix instead of  $V$ .

If there are only a few groups and they are well-separated, and the between-group variation, then projections of the feature data  $y_j$  onto the first few principal axes should depict the group structure. However, a PCA of  $V$  may not always be useful. This point was stressed by Chang (1983), who showed in the case of two groups that the principal component of the feature vector that provides the best separation between the two groups in terms of Mahalanobis distance is not necessarily the first component  $a_1^T y_j$ .



### 3.2 Single-Factor Analysis Model

Geoffrey McLachlan and David Peel (2000) let  $Y_1, \dots, Y_n$  denote a random sample of size  $n$  on a  $p$ -dimensional random vector. In a typical factor analysis model, each observation  $Y_j$  is modeled as:

$$Y_j = \mu + BU_j + e_j \quad (j = 1, \dots, n), \quad (3.1)$$

where  $U_j$  is a  $q$ -dimensional ( $q < p$ ) vector of latent or unobservable variables called factors and  $B$  is a  $p \times q$  matrix of factor loadings (parameters). It is assumed that

$$(Y_1^T, U_1^T)^T, \dots, (Y_n^T, U_n^T)^T$$

are i.i.d. The  $U_j$  are assumed to be i.i.d. as  $N(0, I_q)$ , independently of the errors  $e_j$ , which are assumed to be i.i.d. as  $N(0, D)$ , where  $D$  is a diagonal matrix,

$$D = \text{diag}(\sigma_1^2, \dots, \sigma_p^2),$$

and where  $I_q$  denotes the  $q \times q$  identity matrix. The  $\sigma_i^2$  are called the uniquenesses. Thus, conditional on the  $u_j$ , the  $Y_j$  are independently distributed as  $N(\mu + Bu_j, D)$ . Unconditionally, the  $Y_j$  are i.i.d. according to a normal distribution with mean  $\mu$  and covariance matrix

$$\Sigma = BB^T + D \quad (3.2)$$

Under the model (3.1), the variables  $Y_j$  are conditionally independent given  $u_j$ . Thus the factors in  $u_j$  explain the correlations between the variables in  $Y_j$ , while the error terms  $e_j$  denote the unexplained noise unique to a particular  $y_j$  ( $j = 1, \dots, n$ ). We should mention that in case of  $q > 1$ , there is a plethora of choices for  $B$ , since this model is still satisfied if we replace  $u_j$  by  $Hu_j$  and  $B$  by  $BH^T$ , where  $H$  is any orthogonal matrix of order  $q$ . As  $\frac{1}{2}q(q-1)$  constraints are needed for the unique definition of  $B$ , the number of free parameter is

$$pq + p - \frac{1}{2}q(q-1)$$

If  $q$  is chosen sufficiently smaller than  $p$  so that the difference



$$\begin{aligned}
C &= \frac{1}{2}p(p+1) - pq - p + \frac{1}{2}q(q-1) \\
&= \frac{1}{2}\{(p-q)^2 - (p+q)\} \quad (3.3)
\end{aligned}$$

is positive, then the representation (3.2) imposes some constraints on the covariance matrix  $\Sigma$  and thus reduces the number of free parameters that should be estimated.

### 3.3 EM Algorithm for a Single-Factor Analyzer

McLachlan and Peel (2000) support that the factor analysis model (3.2) can be fitted by maximum likelihood, although we have to compute iteratively the solution as no closed-form expressions exist for the MLEs of  $B$  and  $D$ . The MLE of the mean  $\mu$  is obviously the sample mean  $\bar{y}$  of the  $n$  observed values  $y_1, \dots, y_n$  corresponding to the random sample  $Y_1, \dots, Y_n$ . Hence,  $\mu$  can be replaced by  $\bar{y}$  without loss of generality. Therefore, we let the parameter vector  $\Psi$  of unknown parameters to involve the elements of  $B$  and the diagonal elements of  $D$ . The (incomplete-data) log likelihood for  $\Psi$  that come from the observed data  $y = (y_1^T, \dots, y_n^T)^T$  is, apart from an additive constant,

$$\log L(\Psi) = -\frac{1}{2}n \left\{ \log |BB^T + D| + \sum_{j=1}^m (y_j - \bar{y})^T (BB^T + D)^{-1} (y_j - \bar{y}) \right\}$$

In order to apply the EM algorithm and its variants to this problem, we formulate

$$y_c = (y^T, u_1^T, \dots, u_n^T)^T$$

As the complete-data vector, where  $u_j$  corresponds to  $U_j$ . The complete-data log likelihood is, but for an additive constant,

$$\log L_c(\Psi) = -\frac{1}{2}n \log |D| - \frac{1}{2} \sum_{j=1}^n \left\{ (y_j - \bar{y} - Bu_j)^T D^{-1} (y_j - \bar{y} - Bu_j) + u_j^T u_j \right\}$$

The complete-data density belongs to the exponential family, and the complete-data statistics are  $C_{yy}$ ,  $C_{yu}$  and  $C_{uu}$ , where



$$C_{yy} = \sum_{j=1}^n (y_j - \bar{y})(y_j - \bar{y})^T; C_{yu} = \sum_{j=1}^n (y_j - \bar{y})u_j^T; C_{uu} = \sum_{j=1}^n u_j u_j^T$$

To calculate the conditional expectations of these sufficient statistics given the observed data, we have to use the result that the random vector  $(Y_j^T, U_j^T)^T$  has a multivariate normal distribution with mean  $\begin{pmatrix} \mu \\ 0 \end{pmatrix}$  and covariance matrix  $\begin{pmatrix} BB^T + D & B \\ B^T & I_q \end{pmatrix}$ . As a result, the conditional distribution of  $U_j$  given  $y_j$  is given by

$$U_j | y_j \sim N(\gamma^T (y_j - \mu), I_q - \gamma^T B) \text{ for } j = 1, \dots, n,$$

$$\text{Where, } \gamma = (BB^T + D)^{-1}B.$$

The EM algorithm is performed as follows on the  $(h + 1)$ th iteration.

**E-Step.** Given the current fit  $\Psi^{(h)}$  for  $\Psi$ , calculate as follows the conditional expectation of these sufficient statistics given the observed data  $y$ :

$$E_{\Psi^{(h)}}(C_{yy} | y) = C_{yy}$$

$$E_{\Psi^{(h)}}(C_{yu} | y) = C_{yy}\gamma^{(h)}$$

$$\text{and } E_{\Psi^{(h)}}(C_{uu} | y) = \gamma^{(h)T} C_{yy} \gamma^{(h)} + n w^{(h)}$$

$$\text{where } \gamma^{(h)} = \{B^{(h)}B^{(h)T} + D^{(h)}\}^{-1}B^{(h)}$$

$$\text{and } w^{(h)} = I_q - \gamma^{(h)T} B^{(h)}$$

**M-Step.** Calculate

$$B^{(h+1)} = C_{yy}\gamma^{(h)}(\gamma^{(h)T} C_{yy} \gamma^{(h)} + n w^{(h)})^{-1}$$

And  $D^{(h+1)}$

$$\begin{aligned} &= n^{-1} \text{diag} \left\{ C_{yy} - C_{yy}\gamma^{(h)}(\gamma^{(h)T} C_{yy} \gamma^{(h)} + n w^{(h)})^{-1} \gamma^{(h)T} C_{yy} \right\} \\ &= n^{-1} \text{diag} \left\{ C_{yy} - C_{yy}\gamma^{(h)}B^{(h+1)T} \right\} \end{aligned}$$



The inversion of the current value of the  $p \times p$  matrix  $(BB^T + D)$  on each iteration can be resumed using the result that

$$(BB^T + D)^{-1} = D^{-1} - D^{-1}B(I_q + B^T D^{-1}B)^{-1}B^T D^{-1} \quad (3.4)$$

Where the right hand side of (3.3) involves only the inverses of  $q \times q$  matrices, since  $D$  is a diagonal matrix. The determinant of  $(BB^T + D)$  can be calculated as

$$|BB^T + D| = |D|/|I_q - B^T(BB^T + D)^{-1}B|$$

Liu and Rubin (1994, 1998) considered the application of the ECME algorithm as a solution to this problem. They replaced the M-step by two CM-steps. On the first CM-step  $B^{(h+1)}$  is calculated as on M-step above, while on the second CM-step we obtain the diagonal matrix  $D^{(h+1)}$  by using an algorithm such as Newton-Raphson in order to maximize the actual log likelihood with  $B$  fixed at  $B^{(h+1)}$ .

### 3.4 Mixtures of Factor Analyzers

The scope of its application is limited, because the single-factor analysis model (3.1) provides only a global linear model for the representation of the data in a lower-dimensional subspace. We can obtain a global nonlinear approach by postulating a finite mixture of linear submodels for the distribution of the full observation vector  $Y_j$  given the (unobservable) factors  $u_j$ . That is, we can have a local dimensionality reduction method, as it is mentioned by McLachlan and Peel (2000), by assuming that the distribution of the observation  $Y_j$  can be modeled as

$$Y_j = \mu_i + B_i U_{ij} + e_{ij}, \quad \text{with prob. } \pi_i \quad (i = 1, \dots, g)$$

for  $j = 1, \dots, n$ , where the factors  $U_{i1}, \dots, U_{in}$  are distributed independently  $N(0, I_q)$ , independently of the  $e_{ij}$ , which are distributed independently  $N(0, D_i)$ , where  $D_i$  is a diagonal matrix ( $i = 1, \dots, g$ ). The mixing proportions  $\pi_i$  are nonnegative and sum to one.



Unconditionally, the density of each observation  $Y_j$  is a mixture of  $g$  normal densities in proportions  $\pi_1, \dots, \pi_g$ ; that is,

$$f(y_j; \Psi) = \sum_{i=1}^g \pi_i \varphi(y_j; \mu_i, \Sigma_i) \quad (3.5)$$

$$\text{Where, } \Sigma_i = B_i B_i^T + D_i \quad (i = 1, \dots, g) \quad (3.6)$$

The parameter vector  $\Psi$  now contain the elements of the  $\mu_i$ , the  $B_i$ , and the  $D_i$ , along the mixing proportions  $\pi_i$  ( $i = 1, \dots, g - 1$ ), on putting  $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$ .

The mixtures of factor analyzers model (3.5) is also useful in the modeling of high-dimensional data by mixtures of normal components. With the fitting of a mixture of normal components with unrestricted covariance matrices  $\Sigma_i$ , there are  $\frac{1}{2}p(p+1)$  parameters for each  $\Sigma_i$  ( $i = 1, \dots, g$ ). So, if the number of components  $g$  in the mixture model increases, the result is that the total number of parameters can quickly become very large relative to the sample size  $n$ , which leads to overfitting. The mixture of factor analyzers model gives us the opportunity to control the number of parameters through the reduced model (2.6) for the component-covariance matrices. It thus provides a model intermediate between the independent and unrestricted models. We can easily test the adequacy of the fit of a mixture of factors analyzers with  $q$  factors by using the likelihood ratio test, as McLachlan and Peel (2000) have mentioned.

The number of parameters still might not be manageable, even after the application of MFA approach, particularly if the number of dimensions  $p$  is large and/or the number of components (clusters)  $g$  is not small.

Baek and J. McLachlan (2008) considered how they could modify this factor-analytic approach in order to provide a greater reduction in the number of parameters. They proposed the normal mixture model (3.5) with the bellow restrictions as an extension of model (3.3).

$$\mu_i = A\xi_i \quad (i = 1, \dots, g)$$

And



$$\Sigma_i = A\Omega_i A^T + D \quad (i = 1, \dots, g)$$

Where  $A$  is a  $p \times q$  identity matrix  $\xi_i$  is a  $q$ -dimensional vector,  $\Omega_i$ , is a  $q \times q$  positive definite symmetric matrix, and  $D$  is a diagonal  $p \times p$  matrix. As to be made more precise in the next section,  $A$  is a matrix of loadings on  $q$  unobservable factors and its  $p$  columns are taken to be orthonormal; that is,

$$A^T A = I_q$$

Where  $I_q$  is the  $q \times q$  identity matrix. With these restrictions on the component mean  $\mu_i$  and covariance matrix  $\Sigma_i$ , respectively, the total number of parameters is reduced to

$$d_2 = (g - 1) + p + q(p + g) + \frac{1}{2}(g - 1)q(q + 1).$$

### 3.5 AECM ALGORITHM FOR FITTING MIXTURES OF FACTOR ANALYZERS

#### 3.5.1 AECM Framework

McLachlan and Peel (2000) support that the log likelihood for  $\Psi$  that can be formed from the observed data  $y$  under model (3.5) is

$$\log L(\Psi) = \sum_{j=1}^n \log \left\{ \sum_{i=1}^g \pi_i \varphi(y_j; \mu_i, \Sigma_i) \right\}$$

We can use the alternative expectation-conditional maximization (AECM) algorithm to fit the mixture of factor analyzers model (3.5) by maximum likelihood. The AECM algorithm is an extension of the ECM algorithm, where, on each CM-step, the specification of the complete data is allowed to be different.

To apply the AECM algorithm to the fitting of the mixture model (3.5), we partition the vector of unknown parameters  $\Psi$  as  $(\Psi_1^T, \Psi_2^T)^T$ , where  $\Psi_1$  consists of the mixing proportions  $\pi_i$  ( $i = 1 \dots g - 1$ ) and the elements of the component means  $\mu_i$  ( $i = 1, \dots, g$ ). The subvector  $\Psi_2$  consists of the elements of the  $B_i$  and the  $D_i$  ( $i = 1, \dots, g$ ). Concerning the specification of the incomplete data, it is useful each observation  $y_j$  to be conceptualized as having arisen from one of the components of the mixture and then to declare the component-indicator vector  $z_j$  so associated with  $y_j$  as missing



data. In this framework,  $z_{ij} = (z_j)_i$  is one or zero, according to whether  $y_j$  arose or did not arise from the  $i$ th component ( $i = 1, \dots, g; j = 1, \dots, n$ ). The conditional expectation of  $Z_{ij}$  given  $y_j$  is the posterior probability that the  $j$ th observation comes from the  $i$ th component, given by

$$\begin{aligned}\tau_i(y_j; \Psi) &= pr\{Z_{ij} = 1 | y_j\} \\ &= \frac{\pi_i \varphi(y_j; \mu_i, \Sigma_i)}{\sum_{h=1}^g \pi_h \varphi(y_j; \mu_h, \Sigma_h)}\end{aligned}$$

where  $\Sigma_i$  has the form (2.6) ( $i = 1, \dots, g; j = 1, \dots, n$ ).

We let  $\Psi^{(h)} = (\Psi_1^{(h)T}, \Psi_2^{(h)T})^T$  be the value of  $\Psi$  after  $h$  iterations of the AECM algorithm. For this application of the AECM algorithm, one iteration includes two cycles, and there is one E-step and one CM-step for each cycle. The two CM-steps correspond to the partition of  $\Psi$  into two subvectors  $\Psi_1$  and  $\Psi_2$ .

### 3.5.2 First Cycle

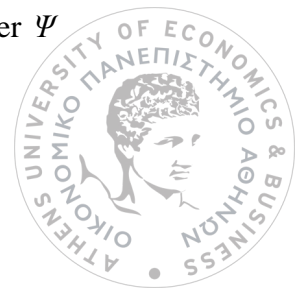
For the first cycle of AECM algorithm, McLachlan and Peel (2000) specify the missing data to be just the component-indicator vectors  $z_1, \dots, z_n$ . The complete-data log likelihood is then given by

$$\log L_c(\Psi) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log\{\pi_i \varphi(y_j; \mu_i, \Sigma_i)\} \quad (3.7)$$

Hence the E-step on the first cycle on the  $(h + 1)$ th iteration requires the calculation of  $Q_1(\Psi; \Psi^{(h)})$ , where

$$Q_1(\Psi; \Psi^{(h)}) = E_{\Psi^{(h)}}\{\log L_c(\Psi) | y\}$$

is the conditional expectation of the complete-data log likelihood (3.7) given  $y$  using  $\Psi^{(h)}$  for  $\Psi$ . This E-step is achieved simply by replacing each  $z_{ij}$  in (3.7) by its current conditional expectation given the observed data (effectively  $y_j$ ); that is, we replace  $z_{ij}$  by  $\tau_i(y_j; \Psi^{(h)})$ . The first CM-step is performed by maximizing  $Q_1(\Psi; \Psi^{(h)})$  over  $\Psi$



with  $\Psi_2$  held fixed at  $\Psi_2^{(h)}$ . The updated estimate  $\Psi_1^{(h+1)}$  of  $\Psi_1$  so obtained contains the new estimates of the  $\pi_i$  and  $\mu_i$  given by

$$\pi_i^{(h+1)} = \sum_{j=1}^n \tau_i(y_j; \Psi^{(h)})/n$$

$$\text{and } \mu_i^{(h+1)} = \sum_{j=1}^n \tau_i(y_j; \Psi^{(h)})y_j / \sum_{j=1}^n \tau_i(y_j; \Psi^{(h)})$$

for  $i = 1, \dots, g$ . We now set  $\Psi^{(h+\frac{1}{2})}$  equal to  $(\Psi_1^{(h+1)T}, \Psi_2^{(h)T})^T$ .

### 3.5.3 Second Cycle

For the second cycle for the updating of  $\Psi_2$ , which consists of the element of the  $B_i$  and the  $D_i$ , McLachlan and Peel (2000) specify the missing data to the factors  $u_1, \dots, u_n$ , as well as the component-indicator vectors  $z_1, \dots, z_n$ . The complete-data log likelihood is now given by

$$\log L_c(\Psi) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \{ \pi_i \varphi(y_j; \mu_i + B_i u_{ij}, \Sigma_i) \}$$

The E-step on the second cycle on the  $(h+1)$ th iteration therefore requires the calculation of  $Q_2(\Psi; \Psi^{(h+\frac{1}{2})})$ , which denotes the conditional of (3.7) given the observed data  $y$ , using  $\Psi^{(h+\frac{1}{2})}$  for  $\Psi$ . In addition in order to update the posterior probabilities of component membership to  $\tau_i(y_j; \Psi^{(h+\frac{1}{2})})$ , it is required the calculation of the conditional expectation of

$$E_{\Psi^{(h+\frac{1}{2})}} \{ Z_{ij} (U_{ij} - \mu_i) | y_j \}$$

and

$$E_{\Psi^{(h+\frac{1}{2})}} \{ Z_{ij} (U_{ij} - \mu_i) (U_{ij} - \mu_i)^T | y_j \}$$

which are given by

$$\tau_i(y_j; \Psi^{(h+\frac{1}{2})}) \gamma_i^{(h)T} (y_i - \mu_i)$$



$$\text{and } \tau_i \left( y_j; \Psi^{(h+\frac{1}{2})} \right) \left\{ \gamma_i^{(h)T} (y_i - \mu_i) (y_i - \mu_i)^T \gamma_i^{(h)} + w_i^{(h)} \right\}$$

$$\text{respectively, where } \gamma_i^{(h)} = \left( B_i^{(h)} B_i^{(h)T} + D_i^{(h)} \right)^{-1} B_i^{(h)}$$

$$\text{and } w_i^{(h)} = I_q - \gamma_i^{(h)} B_i^{(h)} \text{ for } i = 1, \dots, g.$$

The CM-step on this second cycle is implemented by the maximization of  $Q_2 \left( \Psi; \Psi^{(h+\frac{1}{2})} \right)$  over  $\Psi$  with  $\Psi_1$  set equal to  $\Psi_1^{(h+1)}$ . This yields the updated estimates  $B_i^{(h+1)}$  and  $D_i^{(h+1)}$  for  $B_i$  and  $D_i$ , given by

$$B_i^{(h+1)} = V_i^{(h+\frac{1}{2})} \gamma_i^{(h)} \left( \gamma_i^{(h)T} V_i^{(h+\frac{1}{2})} \gamma_i^{(h)} + w_i^{(h+1/2)} \right)^{-1}$$

$$\text{and } D_i^{(h+1)} = \text{diag} \left\{ V_i^{(h+\frac{1}{2})} - V_i^{(h+\frac{1}{2})} \gamma_i^{(h)} B_i^{(h+1)T} \right\}$$

$$\text{where, } V_i^{(h+\frac{1}{2})} = \frac{\sum_{j=1}^n \tau_i \left( y_j; \Psi^{(h+\frac{1}{2})} \right) (y_j - \mu_i^{(h+1)}) (y_j - \mu_i^{(h+1)})^T}{\sum_{j=1}^n \tau_i \left( y_j; \Psi^{(h+\frac{1}{2})} \right)}$$

by construction of this AECM algorithm,

$$Q_1 \left( \Psi^{(h+\frac{1}{2})}; \Psi^{(h)} \right) \geq Q_1 \left( \Psi^{(h)}; \Psi^{(h)} \right)$$

$$\text{And } Q_2 \left( \Psi^{(h+1)}; \Psi^{(h+\frac{1}{2})} \right) \geq Q_2 \left( \Psi^{(h+\frac{1}{2})}; \Psi^{(h+\frac{1}{2})} \right)$$

which ensures that,

$$L \left( \Psi^{(h+\frac{1}{2})} \right) \geq L \left( \Psi^{(h)} \right)$$

and

$$L \left( \Psi^{(h+1)} \right) \geq L \left( \Psi^{(h+\frac{1}{2})} \right)$$



respectively. Thus the (incomplete-data) likelihood  $L(\Psi)$  is not decreased after each cycle and hence after iteration, of the AECM algorithm.

### 3.6 Mixtures of Common Factor Analyzers (MCFA)

Baek and McLachlan (2008) examined the motivation noting the MCFA approach with its constraints  $\mu_i = A\xi_i$  and  $\Sigma_i = A\Omega_i A^T + D$  on the  $g$  component means and covariance matrices  $\mu_i$  and  $\Sigma_i$  ( $i = 1, \dots, g$ ). They tried to show that it can be viewed as a special case of the MFA approach.

The MFA approach with the factor-analytic representation (3.2) on  $\Sigma_i$  is equivalent to assuming that the distribution of the difference  $Y_j - \mu_i$  can be modeled as

$$Y_j - \mu_i = B_i U_{ij} + e_{ij} \text{ with prob. } \pi_i \text{ (} i = 1, \dots, g \text{)} \quad (3.8)$$

for  $j = 1, \dots, n$ , where the (unobservable) factors  $U_{i1}, \dots, U_{in}$  are distributed independently  $N(0, I_q)$ , independently of the  $e_{ij}$ , which are distributed independently  $N(0, D_i)$ , where  $D_i$  is a diagonal matrix ( $i = 1, \dots, g$ ).

As it was mentioned in section 3.4, possibly this model will not lead to a sufficiently large enough reduction in the number of parameters, particularly if  $g$  is not small. If this happen, it is proposed the MCFA approach whereby the distribution of  $Y_j$  is modeled as

$$Y_j = AU_{ij} + e_{ij} \text{ with prob. } \pi_i \text{ (} i = 1, \dots, g \text{)} \quad (3.9)$$

for  $j = 1, \dots, n$  where the (unobservable) factors  $U_{i1}, \dots, U_{in}$  are distributed independently  $N(\xi, \Omega_i)$ , independently of the  $e_{ij}$ , which are distributed independently  $N(0, D)$ , where  $D$  is a diagonal matrix ( $i = 1, \dots, g$ ). Here  $A$  is a  $p \times q$  matrix of factor loadings, which we take to satisfy the relationship:  $A^T A = I_q$

In order to ensure that the MCFA model as specified by (3.9) is a special case of the MFA approach as specified by (3.8), we can rewrite (3.9) as



$$\begin{aligned}
Y_j &= AU_{ij} + e_{ij} = \\
&A\xi_i + A(U_{ij} - \xi_i) + e_{ij} = \\
\mu_i + AK_i K_i^{-1}(U_{ij} - \xi_i) + e_{ij} = \\
\mu_i + B_i + U_{ij}^* + e_{ij} \quad (3.10)
\end{aligned}$$

where  $\mu_i = A\xi_i$

$$B_i = AK_i$$

$$U_{ij}^* = K_i^{-1}(U_{ij} - \xi_i)$$

and where  $U_{ij}^*$  are distributed independently  $N(0, I_q)$ . The covariance matrix of  $U_{ij}^*$  it equal to  $I_q$ , since  $K_i$  can be been chosen so that

$$K_i^{-1}\Omega_i K_i^{-1T} = I_q, \quad (i = 1, \dots, g)$$

On comparing (3.10) with (3.8), we can see that the MCFA model is a special case of the MFA model with the additional restrictions that

$$\begin{aligned}
\mu_i &= A\xi_i \quad (i = 1, \dots, g) \\
B_i &= AK_i \quad (i = 1, \dots, g) \quad (3.11) \\
D_i &= D \quad (i = 1, \dots, g)
\end{aligned}$$

The latter restriction concerning the equal diagonal covariance matrices for the component-specific error terms ( $D_i = D$ ) is sometimes imposed with applications of the MFA approach to avoid potential singularities with small clusters.

Concerning the restriction (3.11) that the matrix of factor of loadings is equal to  $AK_i$  for each component, it could be considered as adopting common factor loadings before the use of the transformation  $K_i$  to transform the factors so that they have zero covariances and unit variances. Hence this is why this approach is called as mixtures of common factor analyzers. It also differs from the MFA approach in that it considers the factor-analytic representation of the observations  $Y_j$  directly, rather than the error terms  $Y_j - \mu_i$ .



As the MFA approach allows a more general representation of the component covariance matrices without any restriction on the component means it is in this sense preferable to the MCFA approach if its application is feasible given the values of  $p$  and  $g$ . If the dimension  $p$  and/or the number of components  $g$  is too large, then the MCFA yields a more feasible approach at the expense of more distributional restrictions on the data. In empirical results we have found the performance of the MCFA approach is usually at least comparable to the MFA approach for data sets to which the latter is practically feasible. The MCFA approach also has the advantage in that the latent factors in its formulation can easily have different means and covariance matrices and are not white noise as happened with the formulation of the MFA approach. Thus the (estimated) posterior means of the factors corresponding to the observed data can be used to depict the latter in low-dimensional spaces.

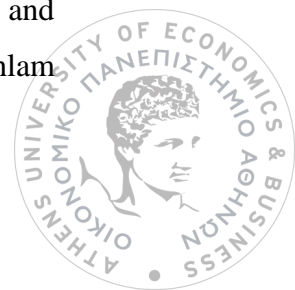
### 3.7 Parsimonious Gaussian Mixture Models

Ghahramani and Hinton (1997) extended the factor analysis model (Section 2.1) by developing the mixture of factor analyzers model which assumes a mixture of Gaussian distributions model with a factor analysis covariance structure for each Gaussian component distribution; this work was further developed by McLachlan and Peel (2000). Additionally, Tipping and Bishop (1999a) developed a mixture of probabilistic principal components model.

Bouveyron and Brunet (2014) claimed that under the general mixture of factor analyzers model, the density of an observation in group  $g$  is of the form with mean parameter  $\mu_i$ , loading matrix  $\Lambda_i$  and noise matrix  $\Psi_i$ . If the probability of membership of group  $g$  considered to be  $\pi_i$ , then this leads to the mixture of factor analyzers model with density

$$f(x_j) = \sum_{i=1}^g \frac{\pi_i}{(2\pi)^{p/2} |\Psi_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x_j - \mu_i - \Lambda_i u_j)' \Psi_i^{-1} (x_j - \mu_i - \Lambda_i u_j) \right\}$$

It would be useful to note that the mixtures of factor analyzers model can be different whether the  $\Psi_i$  term is constrained to be equal across groups or not. Ghahramani and Hinton (1997) assume equal noise and McLachlan and Peel (2000) while McLachlan



et al. (2003) assume unequal noise; however they underline that assuming equal noise can give more stable results. In the context of the mixture of probabilistic principal components analyzers model, Tipping and Bishop (1999a) assume unequal, but isotropic, noise  $\Psi_i = \psi_i I_i$ .

McNicholas and Murphy (2002) proposed extending and unifying these Gaussian mixture models by allowing constraints across groups on the  $\Lambda_i$  and  $\Psi_i$  matrices and on whether or not  $\Psi_i = \psi_i I_i$ . The full range of possible constraints gives a class of eight different parsimonious Gaussian mixture models (PGMM) (Table 3.1).

The Alternating Expectation Conditional Maximization (AECM) (Meng and VanDyk 1997) algorithm is used for fitting these models; McLachlan and Krishnan (1997) also performed the AECM algorithm. This algorithm is an extension of the EM algorithm that uses different definitions of missing data at different stages. For the PGMM, when estimating  $\pi_i$  and  $\mu_i$  the missing data are the unobserved group labels  $z$  and when estimating  $\Lambda_i$  and  $\Psi_i$  the missing data are the group labels  $z$  and the unobserved latent factors  $u$ .

At the first stage of the algorithm, when estimating  $\pi_i$  and  $\mu_i$  we let  $z = (z_1, z_2, \dots, z_n)$  be the group labels of the observations, where

$$z_{ji} = \begin{cases} 1, & \text{if observation } i \text{ belongs to group } g \\ 0, & \text{otherwise} \end{cases}$$

Hence, the complete-data likelihood for the mixture model is

$$L_1(x, z) = \prod_{j=1}^n \prod_{i=1}^g [\pi_i f(x_j | \mu_i, \Lambda_i, \Psi_i)]^{z_{ji}}$$



Model IID	Loading Matrix	Error Variance	Isotopic	Covariance Parameters
CCC	Constrained	Constrained	Constrained	$\{pq - q(q - 1)/2\} + 1$
CCU	Constrained	Constrained	Unconstrained	$\{pq - q(q - 1)/2\} + p$
CUC	Constrained	Unconstrained	Constrained	$\{pq - q(q - 1)/2\} + G$
CUU	Constrained	Unconstrained	Unconstrained	$\{pq - q(q - 1)/2\} + Gp$
UCC	Unconstrained	Constrained	Constrained	$G\{pq - q(q - 1)/2\} + 1$
UCU	Unconstrained	Constrained	Unconstrained	$G\{pq - q(q - 1)/2\} + p$
UUC	Unconstrained	Unconstrained	Constrained	$G\{pq - q(q - 1)/2\} + G$
UUU	Unconstrained	Unconstrained	Unconstrained	$G\{pq - q(q - 1)/2\} + Gp$

Table 3.1: Parsimonious covariance structures derived from the mixture of factor analyzers model.

Hence, the complete-data log-likelihood for the mixture model is

$$\begin{aligned}
 l_1 &= \sum_{j=1}^n \sum_{i=1}^g z_{ji} [\log \pi_i + \log f(x_j | \mu_i, \Lambda_i, \Psi_i)] \\
 &= \sum_{j=1}^n \sum_{i=1}^g z_{ji} \left[ \log \pi_i - \frac{p}{2} \log 2\pi \right. \\
 &\quad \left. - \frac{1}{2} \log |\Lambda_i \Lambda_i' + \Psi_i| - \frac{1}{2} \text{tr} \left\{ (x_j - \mu_i)(x_j - \mu_i)' (\Lambda_i \Lambda_i' + \Psi_i)^{-1} \right\} \right]
 \end{aligned}$$



Hence, we find the expected complete-data log-likelihood is of the form

$$\begin{aligned}
 Q_1(\mu_i, \pi_i) &= \sum_{i=1}^g n_i \log \pi_i - \frac{np}{2} \log 2\pi \\
 &\quad - \sum_{i=1}^g \frac{n_i}{2} \log |\Lambda_i \Lambda_i' + \Psi_i| \\
 &\quad - \sum_{i=1}^g n_i \text{tr} \left\{ \frac{1}{n_i} \sum_{j=1}^n z_{ji} (x_j - \mu_i)(x_j - \mu_i)' (\Lambda_i \Lambda_i' + \Psi_i)^{-1} \right\} \\
 &= \sum_{i=1}^g n_i \log \pi_i - \frac{np}{2} \log 2\pi \\
 &\quad - \sum_{i=1}^g \frac{n_i}{2} \log |\Lambda_i \Lambda_i' + \Psi_i| - \sum_{i=1}^g n_i \text{tr} \{ S_i (\Lambda_i \Lambda_i' + \Psi_i)^{-1} \}
 \end{aligned}$$

Where,  $\hat{z}_{ji} = \frac{\hat{\pi}_i \varphi(x_j | \hat{\mu}_i, \hat{\Lambda}_i, \hat{\Psi}_i)}{\sum_{i'=1}^g \hat{\pi}_{i'} \varphi(x_j | \hat{\mu}_{i'}, \hat{\Lambda}_{i'}, \hat{\Psi}_{i'})}$  (2.12)

$n_i = \sum_{j=1}^n \hat{z}_{ji}$  and  $S_i = \frac{1}{n_i} \sum_{j=1}^n \hat{z}_{ji} (x_j - \mu_i)(x_j - \mu_i)'$ .

Again, the values of  $x$  only appear in this function through  $S_i$ . Maximizing the expected complete-data log-likelihood with respect to  $\pi_i$  and  $\mu_i$  yields

$$\hat{\mu}_i = \frac{\sum_{j=1}^n \hat{z}_{ji} x_j}{\sum_{j=1}^n \hat{z}_{ji}} \text{ and } \hat{\pi}_i = \frac{n_i}{n}$$

At the second stage of the AECM algorithm, when estimating  $\Lambda_i$  and  $\Psi_i$ , we take the group labels  $z$  and the latent factors  $u$  to be the missing data. Therefore, the complete data log likelihood is

$$l_2(x, z, u) = \sum_{j=1}^n \sum_{i=1}^g z_{ji} [\log \pi_i + \log f(x_j | u_j, \mu_i, \Lambda_i, \Psi_i) + \log f(u_j)]$$



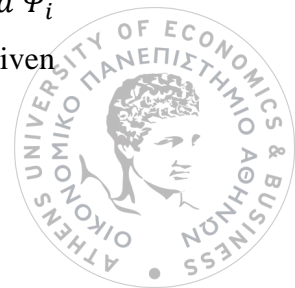
$$\begin{aligned}
&= C + \sum_{i=1}^g \left[ n_i \log \pi_i \right. \\
&\quad - \frac{n_i}{2} \log |\Psi_i| - \frac{n_i}{2} \text{tr} \{ \Psi_i^{-1} S_i \} \\
&\quad \left. + \sum_{j=1}^n z_{ji} (x_j - \mu_i)' \Psi_i^{-1} \Lambda_i u_j - \frac{1}{2} \text{tr} \left\{ \Lambda_i' \Psi_i^{-1} \Lambda_i \sum_{j=1}^n z_{ji} u_j u_j' \right\} \right]
\end{aligned}$$

where  $C$  is a constant function of  $\mu_i$ ,  $\Lambda_i$  and  $\Psi_i$ .

It follows that the expected complete-data log-likelihood evaluated with  $\mu_i = \hat{\mu}_i$  and  $\pi_i = \hat{\pi}_i$  is of the form

$$\begin{aligned}
Q(\Lambda_i, \Psi_i) &= C + \sum_{i=1}^g \left[ -\frac{n_i}{2} \log |\Psi_i| - \frac{n_i}{2} \text{tr} \{ \Psi_i^{-1} S_i \} \right. \\
&\quad \left. + \sum_{j=1}^n \hat{z}_{ji} (x_j - \hat{\mu}_i)' \Psi_i^{-1} \Lambda_i E[u_j | x_j, \hat{\mu}_i, \hat{\Lambda}_i, \hat{\Psi}_i] \right] \\
&= C + \sum_{i=1}^g \left[ -\frac{n_i}{2} \log |\Psi_i| - \frac{n_i}{2} \text{tr} \{ \Psi_i^{-1} S_i \} \right. \\
&\quad + \sum_{j=1}^n \hat{z}_{ji} (x_j - \hat{\mu}_i)' \Psi_i^{-1} \Lambda_i \hat{\beta}_i (x_j - \hat{\mu}_i) \\
&\quad \left. - \frac{1}{2} \text{tr} \left\{ \Lambda_i' \Psi_i^{-1} \Lambda_i \sum_{j=1}^n \hat{z}_{ji} E[u_j u_j' | x_j, \hat{\mu}_i, \hat{\Lambda}_i, \hat{\Psi}_i] \right\} \right] \\
&= C + \sum_{i=1}^g n_i \left[ \frac{1}{2} \log |\Psi_i^{-1}| - \frac{1}{2} \text{tr} \{ \Psi_i^{-1} S_i \} + \text{tr} \{ \Psi_i^{-1} \Lambda_i \hat{\beta}_i S_i \} - \frac{1}{2} \text{tr} \{ \Lambda_i' \Psi_i^{-1} \Lambda_i \Theta_i \} \right]
\end{aligned}$$

where  $\Theta_i = (I_i - \hat{\beta}_i \hat{\Lambda}_i + \hat{\beta}_i S_i \hat{\beta}_i')$  is a symmetric  $q \times q$  matrix and the  $\hat{z}_{ji}$  are computed as in (2.12) with the estimates  $\hat{\mu}_i$  and  $\hat{\pi}_i$  as calculated in the first stage of the AECM algorithm. When we impose constraints (Table 1) on the  $\Lambda_i$  and  $\Psi_i$  matrices, the results can be easily derived from the expression for  $Q(\Lambda_i, \Psi_i)$  given



above. The first stage of the AECM where  $\pi_i$  and  $\mu_i$  are estimated and the second stage where  $\Lambda_i$  and  $\Psi_i$  are estimated and iterated until convergence. The resulting values give maximum likelihood estimates of the parameters in the model. The resulting  $\hat{z}_{ji}$  values are the estimates of the *a posteriori* probability of group membership for each observation.



## Chapter 4

### Application

The analysis of gene expression microarray data using clustering techniques has received the attention of researchers in the discovery, validation, as it has an important role to play in understanding of various classes and subclasses of cancer as mentioned McLachlan et al. (2001). The package we are going to present here, called MCFA, can be applied to the problem of clustering tissue samples on the basis of genes and to the problem of clustering genes on the basis of tissues. The tissue space and the gene space generally differ in dimensionality (10 – 100 tissues versus 1000 – 10000 genes). A standard cluster analysis problem is the clustering of the genes on the basis of the tissues that can be effected by fitting normal mixture models. The genes are assumed to be uncorrelated within a cluster, as without this assumption the clustering of the tissue samples on the basis of all genes is nonstandard since the dimension of each tissue sample (the number of genes) is so much greater than the number of tissues. We can handle the dimensionality problem with the MCFA package, by fitting mixtures of factor analyzers, which allow for nonzero component-correlations between the genes. Given the very large number of genes in a typical tissue sample, we can achieve a reduction in the number of genes to be used in the clustering process.

The MCFA package is to be applied in the clustering of a well-known data set in the microarray literature, the leukaemia data analyzed in Golub *et al.* (1999).

#### 4.1 Dimension Reduction

In the standard setting of a model-based cluster analysis the  $n$  observations to be clustered are taken to be independent realizations where the sample size  $n$  is much larger than the dimension  $p$  of each observation,

$$n \gg p \quad (4.1)$$

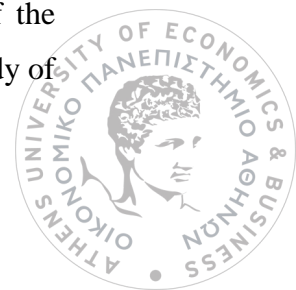


it is also assumed that the sizes of the clusters to be produced are sufficiently large relative to  $p$  in order to avoid any singular estimates of the within-cluster covariance matrices.

We now consider the cluster analysis of microarray data collected on  $n$  genes from  $p$  experiments, which can be represented in the form of a  $n \times p$  data matrix  $A$  whose  $i$ th row contains the expression levels for the  $i$ th gene in the  $p$  tissue samples. As we can see,  $n$  is typically larger than  $p$ . Thus for the problem of clustering  $n$  genes on the basis of the  $p$  tissues are available, and so the condition (4.1) for the standard cluster analysis will be satisfied. Unless all the genes in a given tissue sample are independently distributed, the condition of independent data will not hold. But in practice we can proceed with the standard clustering methodology, ignoring any correlations between genes in the same sample.

We now consider the problem of clustering the  $p$  tissues on the basis of the  $n$  genes. In our case, the dimension  $p$  will be typically large relative to the sample size  $n$ , thus causing estimation problems with the normal mixture model. This is because the  $g$ -component normal mixture model with unrestricted component-covariance matrices is a highly parameterized model with  $\frac{1}{2}p(p+1)$  parameters for each component-covariance matrix  $\Sigma_i$  ( $i = 1, \dots, g$ ). It therefore cannot be fitted directly to the tissues on the basis of all the  $p = n$  genes. The MCFA package handles this high-dimensional problem by using mixtures of factor analyzers, where  $\Sigma_i$  is specified by (2.4) and  $D_i = D$  ( $i = 1, \dots, g$ ). We can achieve a reduction in the number of parameters by taking the number of the factors  $q$  to be appropriately small. Although the model under  $D_i = D$  can be fitted if  $q$  is considered to be less than the sample size  $n$ ,  $q$  needs to be sufficiently small to ensure that the estimates of the component-covariance matrices are not highly variable. Hence  $q$  may not be able always to be taken sufficiently large to model adequately the full correlation structure of the genes in the lower  $q$ -dimensional factor space.

Thus, practically, we may wish to work with a subset of the available genes, particularly as the fitting of a mixture of factor analyzers will involve a considerable amount of computation time for an extremely large number of genes. What is more, the intent of the cluster analysis may not be the production of a clustering of the tissues on the basis of all the available genes, but rather the discovery and the study of



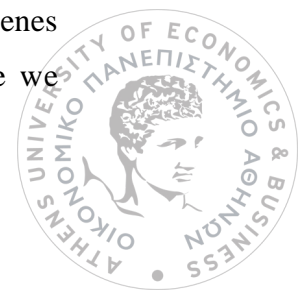
different clusterings of the tissues corresponding to different subsets of the genes. Indeed, the simultaneous use of too many genes in the cluster analysis may serve only to create noise that masks the effect of a smaller number of genes.

There are two optional stages before the final stage of clustering the tissues. According to mentioned McLachlan et al. (2001), the first stage has to do with the selection of a subset of relevant genes from the available set of genes. The second stage then is about the grouping of the retained set of genes into a specified number ( $N_0$ ) of groups. The third and final stage concerns the clustering of the tissues by fitting mixtures of factor analyzers. It can be undertaken on the basis of (i) all or a selected subset of the available genes, (ii) all or some of the gene group means, or (iii) all or some of the genes within a specified gene group.

Some of the subspace clustering approaches, model the data in a low and common subspace which allows the data to be displayed in a low-dimensional plot. The *mca* function for R, which implements the method of Baek et al. (2009), is one of them. It enables model-based density estimation to be undertaken for high-dimensional data, where the dimension is not very small relative to the number of observations. Still for the R software, the *hmfa* function by C. Viroli (2010) implements the approach proposed in Montanari and Viroli (2010), also based on the mixture of factor analyzers.

We are about to apply the MCFA package to the clustering of the leukaemia tissues of Golub *et al.* (1999), as did mentioned McLachlan et al. (2001), who studied gene expressions on two types leukaemias: Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML). Gene expression levels were measured using Affymetrix high density oligonucleotide arrays containing  $n = 7129$  genes on  $p = 72$  tissues, comprising 47 cases of ALL (38 B-cell and 9 T-cell ALL) and 25 cases of AML. McLachlan et al. (2001) followed the processing steps of Dudoit *et al.* (2001) of: (i) filtering: exclusion of gene with  $\max/\min \leq 5$  and  $(\max - \min) \leq 500$ , where  $\max$  and  $\min$  refer respectively to the maximum and minimum expression levels of a particular gene across a tissue sample; (ii) the natural logarithm of the expression levels was taken (Dudoit *et al.*, (2001), used base 10 logarithms). This left us 3731 genes.

McLachlan et al. (2001) reduced this set further to 2015 genes by eliminating genes not considered to be relevant on the first stage of approach. On second stage we



summarized the expression levels on these 2015 selected genes by clustering them into a number of groups ( $N_0 = 40$ ). It was found that Groups 1 and 3 provide clusterings that are most similar to the external classification of the tissues. It is confirmed by fitting a two-component mixture factor analyzer with  $q=6$  factors to the tissues on the basis of the genes in Groups 1-3, respectively. In sequel, we will choose models according to BIC criterion, and it will be ensured that the model of a two or three-component mixture factor analyzer with  $q=6$  is more reliable relative to the others with less factors.

#### 4.2 Choosing models according BIC criterion

As we have mentioned in section 3.3, In order to apply the EM algorithm, on the E-step, given the current fit  $\Psi^{(h)}$  for  $\Psi$ , calculate the conditional expectation of these sufficient statistics given the observed data  $y$ . on the M-step we calculate  $B^{(h+1)}$  and  $D^{(h+1)}$ . The EM algorithm is implemented as follows on the  $(h + 1)$ th iteration.

We have done this procedure by using the package `mcfa`, for relevant numbers of factors and groups. As we can see from the below results, according to BIC criterion, models with more factors are the most reliable.

	q=2	q=3	q=4	q=5	q=6	q=7	q=8
g=2	-46520.48	-44901.66	-43604.42	-42481.33	-41406.33	-40519.57	-39767.96
g=3	-46509.63	-44886.74	-43568.63	-42428.06	-41343.31	-40443.90	-39677.84
g=4	-46502.23	-44874.34	-43550.52	-42396.31	-41298.51	-40386.95	-39616.95
g=5	-46496.33	-44862.60	-43532.01	-42373.94	-41269.27	-40366.08	-39567.12
g=6	-46492.26	-44850.37	-43514.49	-42367.84	-41244.97	-40329.01	-39594.02

**Table 4.1:** Max-likelihood for models according to number of factors and number of groups



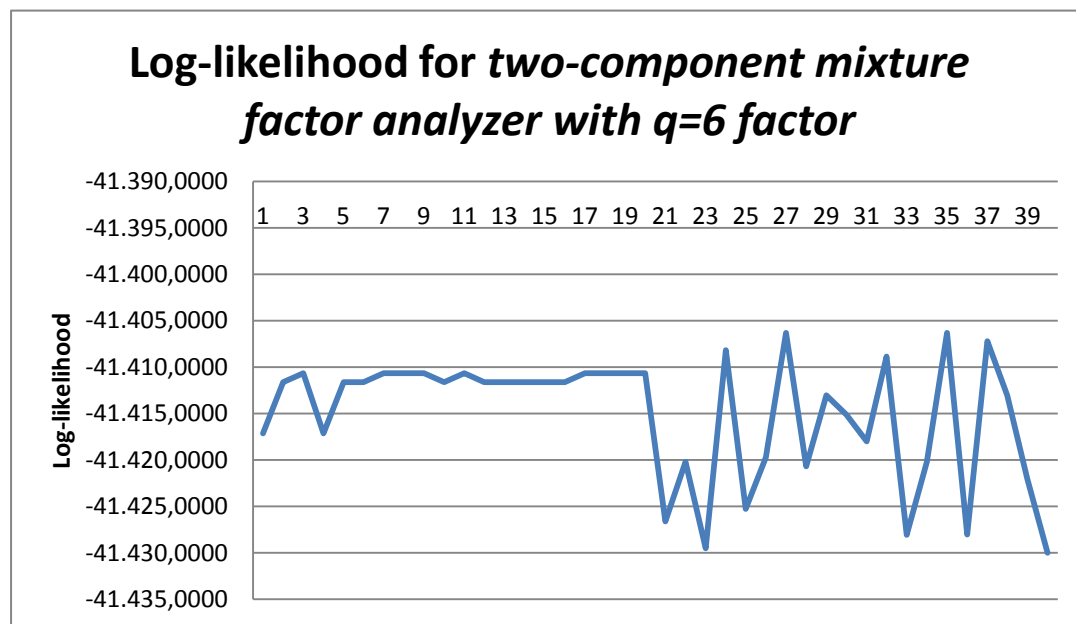
	q=2	q=3	q=4	q=5	q=6	q=7	q=8
g=2	93969.62	90731.98	88137.51	85891.33	83741.33	81967.80	80464.59
g=3	93947.93	90702.15	88065.92	85784.79	83615.29	81816.47	80284.35
g=4	93933.12	90677.34	88029.70	85721.29	83525.69	81702.57	80162.56
g=5	93921.32	90653.87	87992.70	85676.55	83467.20	81660.83	80062.90
g=6	93913.18	90629.41	87957.64	85664.34	83418.60	81586.69	80116.70

**Table 4.2:** BIC for models according to number of factors and number of groups

It was found that models with  $q=6$ ,  $q=7$  and  $q=8$  factors are more reliable from the others with less number of factors, as we can see from table 3.2, according BIC criterion.

#### 4.3 Application of a two-component mixture factor analyzer with $q=6$ factor

Now we will apply the EM algorithm for a fitting a two-component mixture factor analyzer with  $q=6$  factors to the tissues. After 40 iterations we have the log-likelihood=-41430.0165 and maximum log-likelihood = -41406.3344. Above we present the iterations of the application of MCFA package.



**Figure 4.1:** Log-likelihood for a two-component mixture factor analyzer with  $q=6$  factor



As we mentioned in section 3.6, we can assume that the distribution of the observation  $Y_j$  can be modeled as

$$Y_j - \mu_i = +B_i U_{ij} + e_{ij}, \quad (4.1) \quad \text{With prob. } \pi_i \quad (i = 1, \dots, g)$$

for  $j = 1, \dots, n$

by fitting a two-component mixture factor analyzer with  $q=6$  factors to the tissues, we have the below probabilities and the mixing proportions  $\pi_i$  are nonnegative and sum to one.

$\pi_{i1}$	$\pi_{i2}$
0.506	0.494

**Table 4.3:** Component probabilities for dataset with 500 variables, for a two-component mixture factor analyzer with  $q=6$  factors

In (4.1) the factors  $U_{i1}, \dots, U_{in}$  are distributed independently  $N(\xi, \Omega_i)$ , independently of the  $e_{ij}$ . In our application we have:

	Matrix of factor mean vectors	
$U_{i1}$	5.41282	5.5485
$U_{i2}$	1.7648	-1.8090
$U_{i3}$	0.2955	-0.3029
$U_{i4}$	-1.4936	1.5310
$U_{i5}$	0.1924	-0.1972
$U_{i6}$	-1.0067	1.0320

**Table 4.4:** Factor loading matrix for dataset with 500 variables, for a two-component mixture factor analyzer with  $q=6$  factors



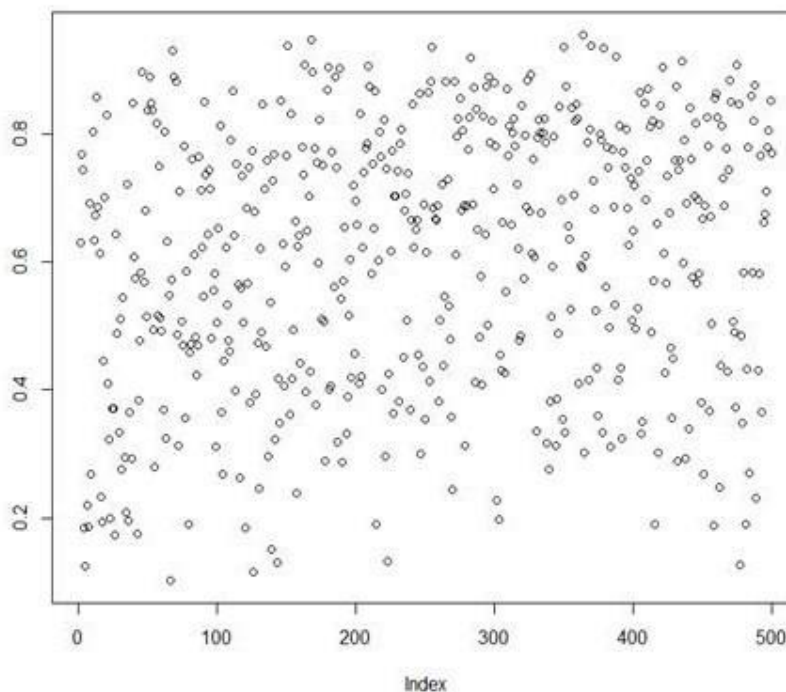
We extended our model with the restrictions:  $\mu_i = A\xi_i$  and  $\Sigma_i = A\Omega_iA^T + D$ , where  $\Omega_i$  is a  $q \times q$  positive definite symmetric matrix.

	Components					
$w_1$	27.4325	20.4768	11.4412	-2.7358	-12.6755	-6.8686
	20.4768	33.5462	8.3148	-0.8006	-10.0460	6.1559
	11.4412	8.3148	12.6546	1.6973	-8.9448	-9.6786
	-2.7358	-0.8006	1.6973	33.0618	7.5471	-4.1928
	-12.6755	-10.0460	-8.9448	7.5471	17.7274	5.3828
	-6.8686	6.1559	-9.6786	-4.1928	5.3828	17.9237
$w_2$	42.7918	-1.3293	-8.1048	-13.4980	14.3165	-3.8270
	-1.3293	29.9186	-9.5158	6.3435	8.5965	-1.8512
	-8.1048	-9.5158	48.9702	-0.9686	9.1186	11.03989
	-13.4980	6.3435	-0.9686	11.4701	-6.8253	0.8334
	14.3165	8.5965	9.1186	-6.8253	21.110	-4.7852
	-3.8270	-1.8512	11.0398	0.8334	-4.7852	12.9533

**Table 4.5:** Array of factor covariance matrix for dataset with 500 variables, for a two-component mixture factor analyzer with  $q=6$  factors.

$U_{ij}$  are independently of the  $e_{ij}$ , which are distributed independently  $N(0, D)$ , where  $D$  is a diagonal matrix as it is represented above.

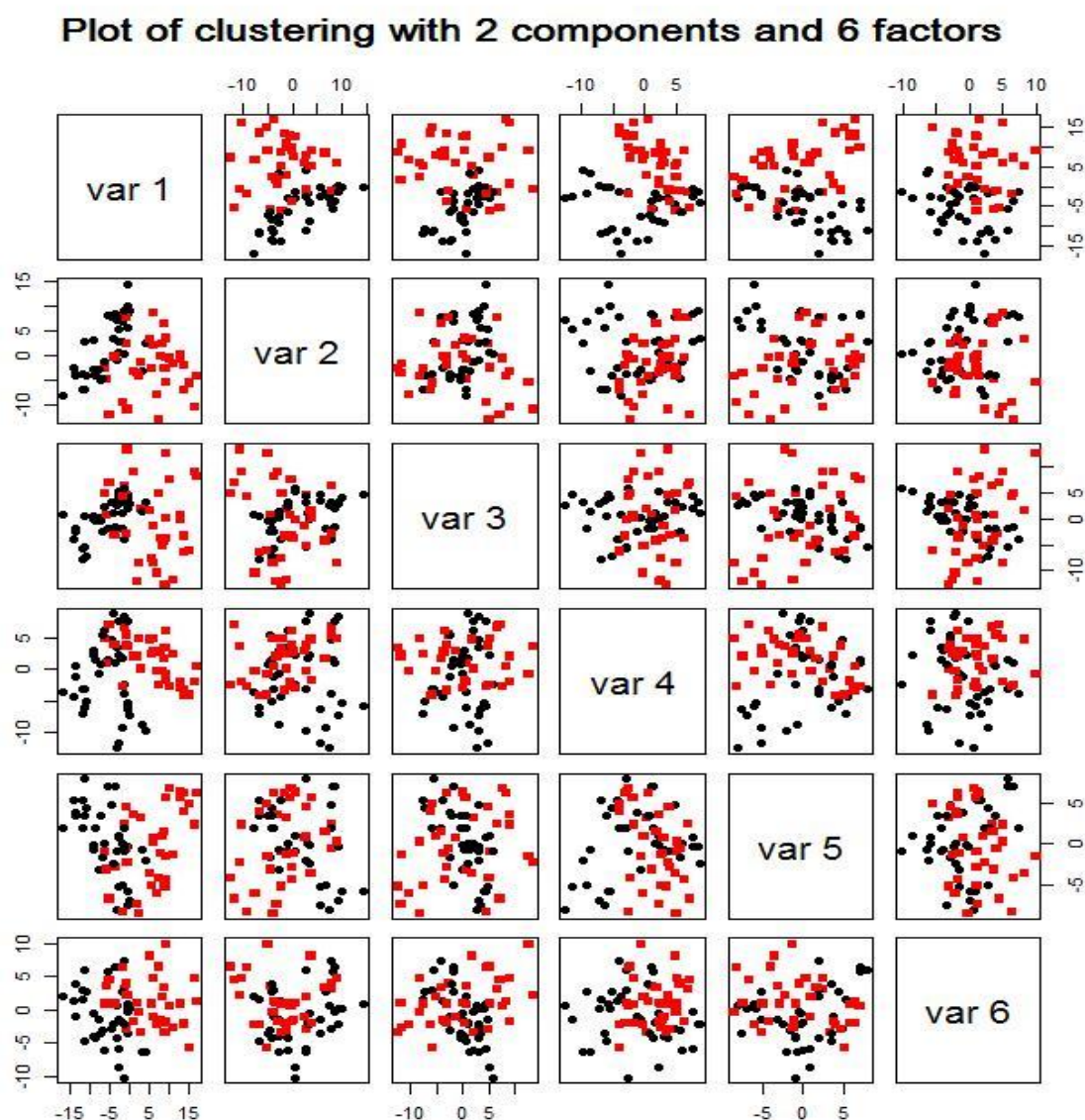
**Plot of data of diagonal matrix D for clustering with 2 components and 6 factors**



**Figure 4.2:** Plot of error covariance matrix for clustering with 2 components and 6 factors.

In Figure 4.3 we plotted the estimated factor scores  $\hat{u}_j = (\hat{u}_{1j}, \hat{u}_{2j}, \hat{u}_{3j}, \hat{u}_{4j}, \hat{u}_{5j}, \hat{u}_{6j})^T$  of the tissues in two-dimensional space according to their clustered membership determined by the MCFA with  $q = 6$  factors. The probability of first group is 0.506, while the probability of the second is 0.494. It can be seen from Figure 3.4 that the two classes ALL and AML are well separated.

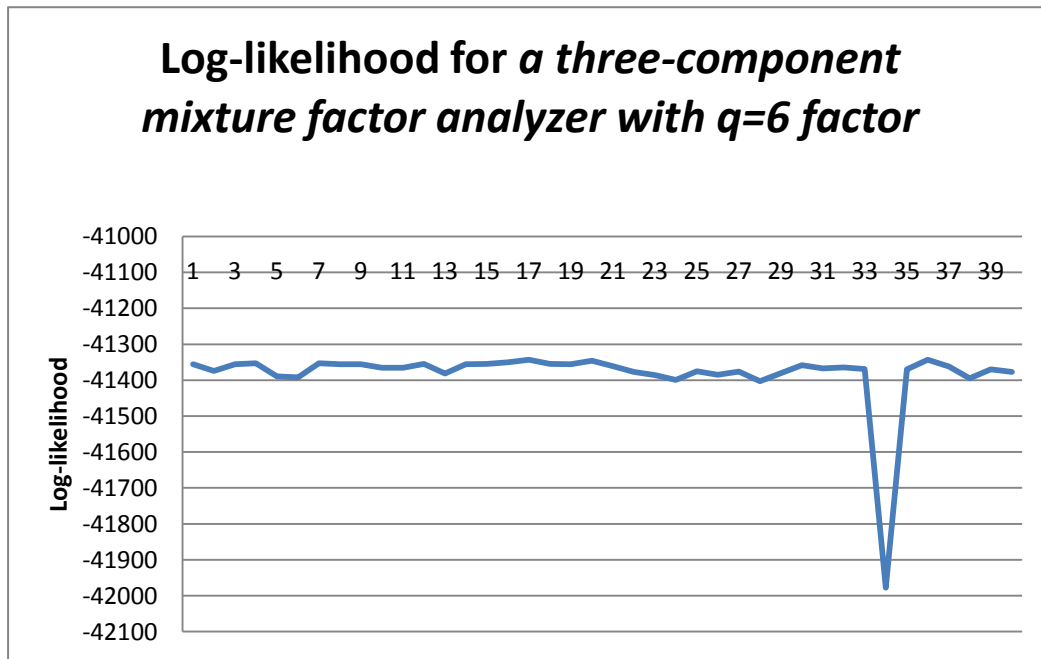
And there is also presented the correlation between the 6 factors.



**Figure 4.3:** Plot of clustering with 2 components and 6 factors.

#### 4.4 Application of a three-component mixture factor analyzer with $q=6$ factor

Now we will apply the EM algorithm for a fitting a three-component mixture factor analyzer with  $q=6$  factors to the tissues. After 40 iterations we have the log-likelihood=-41377.2056 and maximum log-likelihood = -41343.3139. Above we present the iterations of the application of MCFA package.



**Figure 4.4:** Log-likelihood for a three-component mixture factor analyzer with  $q=6$  factor

By fitting in 4.1 a three-component mixture factor analyzer with  $q=6$  factors to the tissues, we have the below probabilities and the mixing proportions  $\pi_i$  are nonnegative and sum to one.

$p_{i1}$	$p_{i2}$	$p_{i3}$
0.166	0.347	0.478

**Table 4.6:** Component probabilities for dataset with 500 variables, for a three-component mixture factor analyzer with  $q=6$  factors.



In (4.1) the factors  $U_{i1}, \dots, U_{in}$  are distributed independently  $N(\xi, \Omega_i)$ , independently of the  $e_{ij}$ . In our application we have:

Matrix of factor mean vectors			
$U_{i1}$	0.2086	8.9064	-6.4233
$U_{i2}$	6.6827	-1.9792	-0.8621
$U_{i3}$	1.2911	-2.1977	1.1252
$U_{i4}$	-6.2491	0.5939	1.7028
$U_{i5}$	-2.8836	-0.4213	1.2833
$U_{i6}$	-0.5906	0.9718	-0.4953

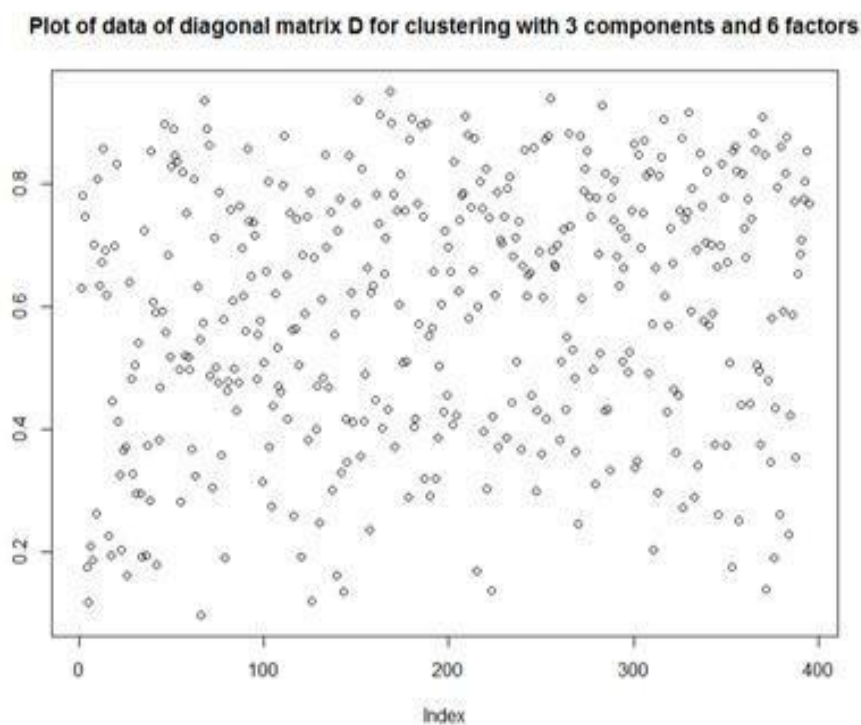
**Table 4.7:** Factor loading matrix for dataset with 500 variables, for a three-component mixture factor analyzer with  $q=6$  factors.

We extended our model with the restrictions:  $\mu_i = A\xi_i$  and  $\Sigma_i = A\Omega_iA^T + D$ , where  $\Omega_i$  is a  $q \times q$  positive definite symmetric matrix.

	Components					
$w_1$	4.6077	-7.3137	1.1183	1.2429	2.999	-6.1028
	-7.3137	24.0285	0.7884	-0.8411	-8.095	11.8624
	1.1183	0.7884	7.70174	-5.8820	-5.544	-7.7082
	1.2429	-0.8411	-5.88204	23.0525	10.503	6.7585
	2.9991	-8.0959	-5.5445	10.5033	11.316	2.3676
	-6.1028	11.8624	-7.7082	6.7585	2.3676	15.562
$w_2$	19.5628	-3.8830	9.2477	-5.4793	17.687	-6.1642
	-3.8830	23.155	-8.1205	9.0593	6.707	-0.1171
	9.2477	-8.1205	47.3916	-11.2662	11.387	15.4297
	-5.4793	9.0593	-11.2662	10.8953	-7.142	-0.2527
	17.6872	6.7079	11.3873	-7.14268	26.057	-5.4861
	-6.1642	-0.1171	15.4297	-0.2527	-5.4861	15.9985
$w_3$	20.0342	11.461	14.0849	13.490	-4.000	-4.9033
	11.4611	28.074	0.4223	11.132	7.651	-1.9979
	14.0849	0.422	21.2396	11.123	-7.305	-5.1467
	13.4909	11.132	11.1233	18.690	-6.377	-2.6945
	-4.0006	7.6515	-7.3050	-6.377	13.192	3.4066
	-4.9033	-1.9979	-5.1467	-2.694	3.406	16.1836

**Table 4.8:** Array of factor covariance matrix for dataset with 500 variables, for a three-component mixture factor analyzer with  $q=6$  factors.

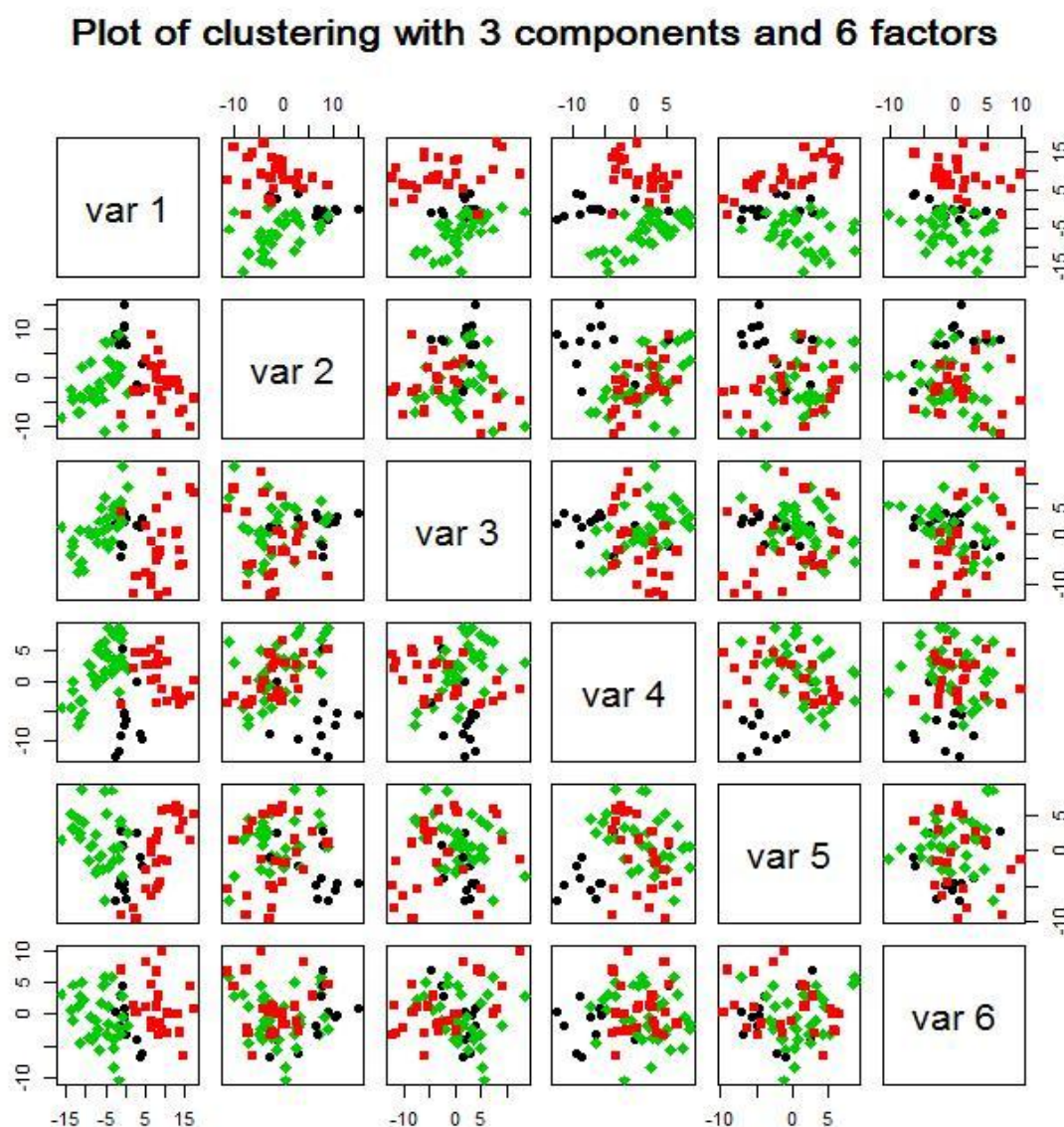
$U_{ij}$  are independently of the  $e_{ij}$ , which are distributed independently  $N(0, D)$ , where  $D$  is a diagonal matrix as it is represented above.



**Figure 4.5:** Plot of error covariance matrix for clustering with 3 components and 6 factors, for dataset with 500 variables.

In Figure 4.6 we plotted the estimated factor scores  $\hat{u}_j = (\hat{u}_{1j}, \hat{u}_{2j}, \hat{u}_{3j}, \hat{u}_{4j}, \hat{u}_{5j}, \hat{u}_{6j})^T$  of the tissues in three-dimensional space according to their clustered membership determined by the MCFA with  $q = 6$  factors. The probability of first group is 0.166, the probability of the second group is 0.347, while the probability of the third one is 0.478. It can be seen from Figure 3.8 that the three classes T-cell ALL, B-cell ALL and AML are well separated.

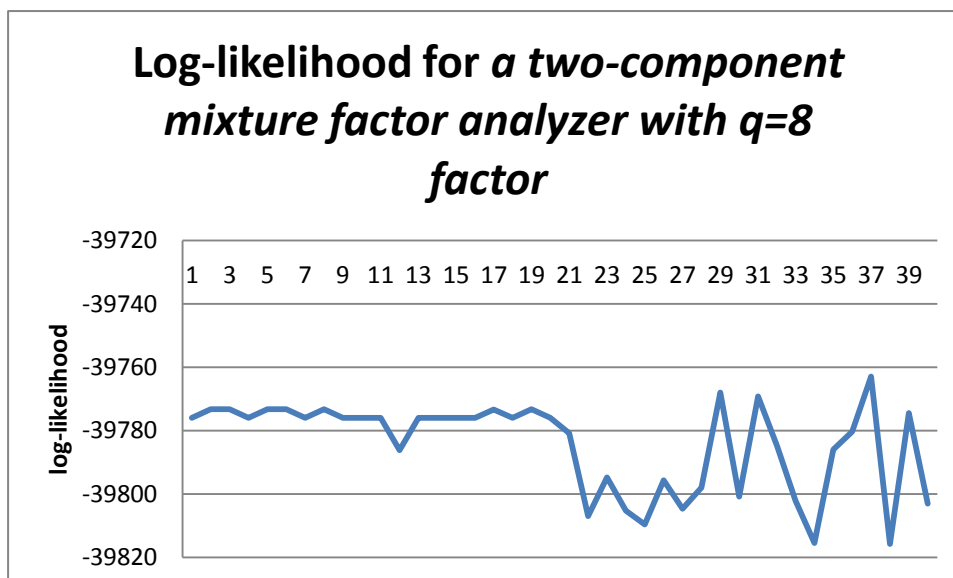
And there is also presented the correlation between the 6 factors.



**Figure 2:** Plot of clustering with 3 components and 6 factors

#### 4.5 Application of a two-component mixture factor analyzer with $q=8$ factor

Now we will apply the EM algorithm for a fitting a two-component mixture factor analyzer with  $q=8$  factors to the tissues. After 40 iterations we have the log-likelihood = -39803.0082 and maximum log-likelihood = -39767.9475. Above we present the iterations of the application of MCFA package.



**Figure 4.7:** Log-likelihood for a two-component mixture factor analyzer with  $q=8$  factor

By fitting in 4.1 a two-component mixture factor analyzer with  $q=8$  factors to the tissues, we have the below probabilities and the mixing proportions  $\pi_i$  are nonnegative and sum to one.

$p_{i1}$	$p_{i2}$
0.481	0.519

**Table 4.9:** Component probabilities, for dataset with 500 variables, for a two-component mixture factor analyzer with  $q=8$  factors.

In (4.1) the factors  $U_{i1}, \dots, U_{in}$  are distributed independently  $N(\xi, \Omega_i)$ , independently of the  $e_{ij}$ . In our application we have:

	Matrix of factor mean vectors	
$U_{i1}$	-5.1952	4.8152
$U_{i2}$	1.8136	-1.6811
$U_{i3}$	0.7880	-0.7299
$U_{i4}$	-1.5565	1.4426
$U_{i5}$	0.36558	-0.3386
$U_{i6}$	-0.7038	0.6524
$U_{i7}$	-1.5550	1.4414
$U_{i8}$	-0.07541	0.0714

**Table 4.10:** Factor loading matrix for dataset with 500 variables, for dataset with 500 variables, for a two-component mixture factor analyzer with  $q=8$  factors.

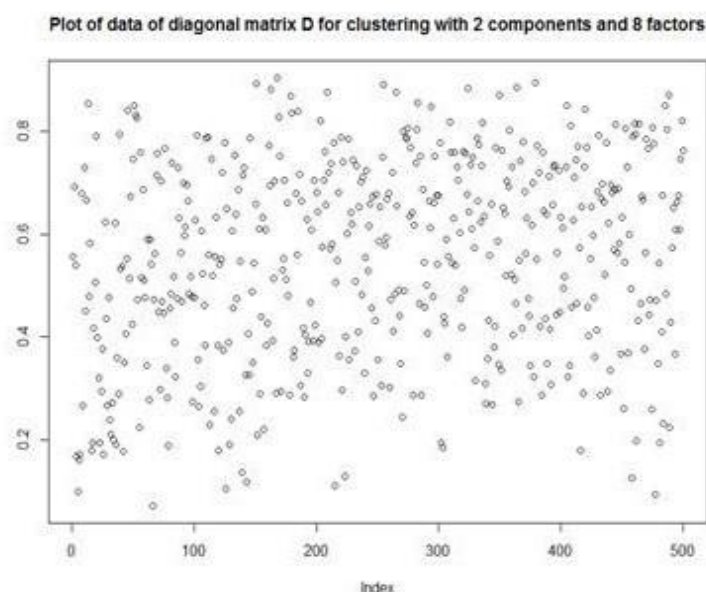


We extended our model with the restrictions:  $\mu_i = A\xi_i$  and  $\Sigma_i = A\Omega_iA^T + D$ , where  $\Omega_i$  is a  $q \times q$  positive definite symmetric matrix.

	Components							
$w_1$	30.7281	20.3772	13.077	-0.8501	-13.6983	-5.354	-8.3022	-1.4904
	20.3772	33.6109	9.696	0.3157	-11.6949	8.131	-3.4132	2.1331
	13.0774	9.6965	11.660	3.0784	-8.7603	-8.320	-2.2292	-1.2248
	-0.8501	0.3157	3.0784	31.6729	9.4667	-1.515	-5.6253	-1.9258
	-13.6983	-11.6949	-8.7603	9.4667	19.0062	5.040	1.3127	-0.9125
	-5.3543	8.1316	-8.3203	-1.5153	5.0405	22.618	-2.5140	2.4080
	-8.3022	-3.4132	-2.2292	-5.6253	1.3127	-2.514	9.9200	0.2878
	-1.4904	2.1331	-1.2248	-1.9258	-0.9125	2.408	0.2878	1.1660
$w_2$	48.7493	-2.0724	-4.9676	-14.0293	16.0057	-1.2406	-6.9073	-0.3145
	-2.0724	30.5013	-11.4857	4.8625	8.9773	-5.0423	7.7416	-2.3447
	-4.9676	-11.4857	47.4267	-0.4754	7.4368	8.1613	4.7441	2.2510
	-14.0293	4.8625	-0.4754	14.1284	-7.3846	-0.5501	1.3163	1.8617
	16.0057	8.9773	7.4368	-7.3846	19.6470	-4.1616	-0.3940	0.4300
	-1.2406	-5.0423	8.1613	-0.5501	-4.1616	9.9815	0.6072	-1.9102
	-6.9073	7.7416	4.7441	1.3163	-0.3940	0.6072	17.6454	-0.1454
	-0.3145	-2.3447	2.2510	1.8617	0.4300	-1.9102	-0.1454	21.6593

**Table 4.11:** Array of factor covariance matrix for dataset with 500 variables, for a two-component mixture factor analyzer with  $q=8$  factors.

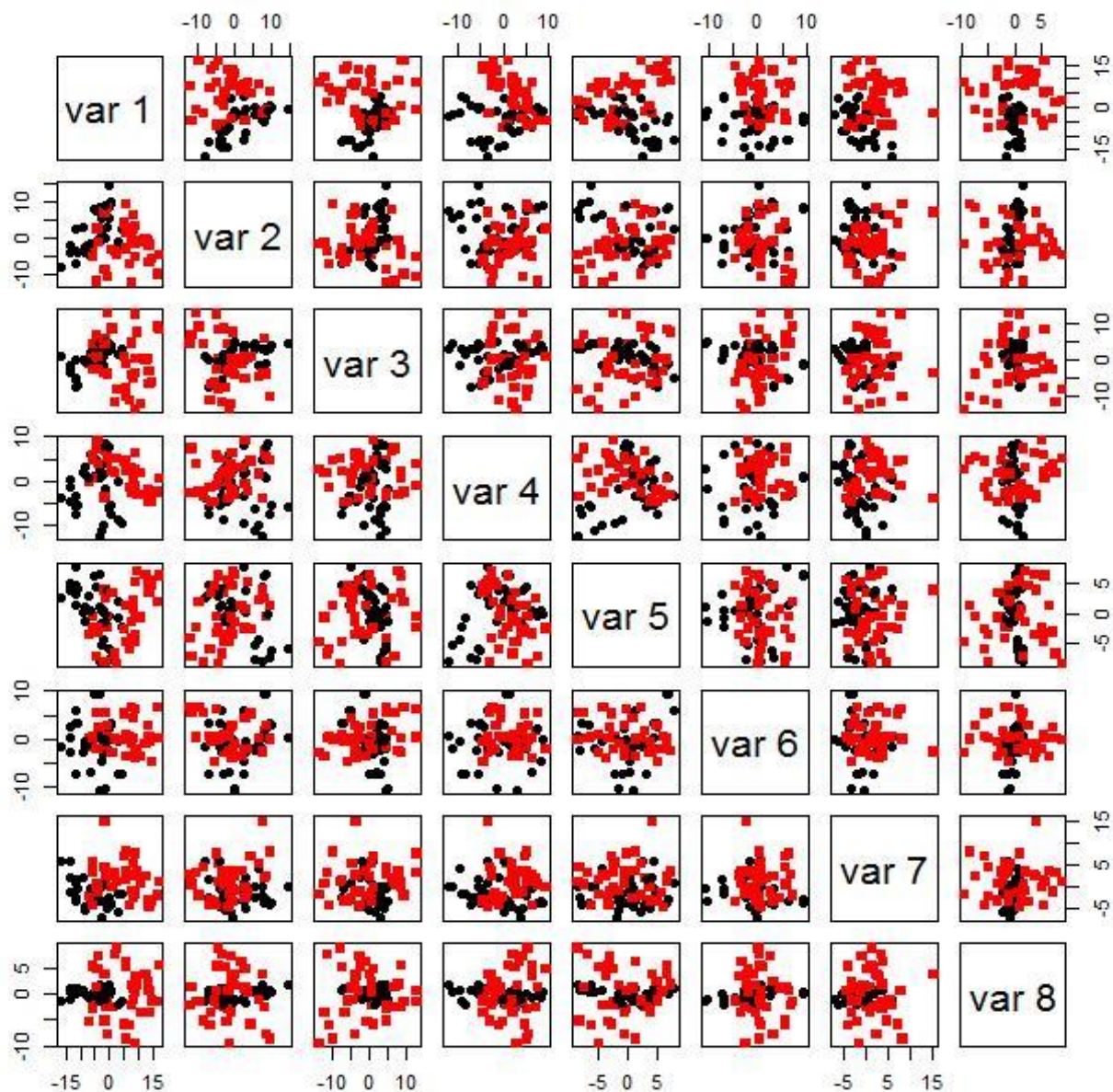
$U_{ij}$  are independently of the  $e_{ij}$ , which are distributed independently  $N(0, D)$ , where  $D$  is a diagonal matrix as it is represented above.



**Figure 4.8:** Plot of error covariance matrix for clustering with 2 components and 6 factors, for dataset with 500 variables.

In Figure 4.9 we plotted the estimated factor scores  $\hat{u}_j = (\hat{u}_{1j}, \hat{u}_{2j}, \hat{u}_{3j}, \hat{u}_{4j}, \hat{u}_{5j}, \hat{u}_{6j}, \hat{u}_{7j}, \hat{u}_{8j})^T$  of the tissues in two-dimensional space according to their clustered membership determined by the MCFA with  $q = 8$  factors. The probability of first group is 0.481, while the probability of the second is 0.519. It can be seen from Figure 3.6 that the two classes ALL and AML are well separated. And there is also presented the correlation between the 8 factors.

### Plot of clustering with 2 components and 8 factors



**Figure 3:** Plot of clustering with 2 components and 8 factors

## 5 Conclusion

Baek and McLachlan (2008) claimed that in practice, the use of normal mixture models in density estimation and clustering received much attention. However, for high-dimensional data sets, the component covariance matrices are highly parameterized and some form of reduction in the number of parameters is needed, particularly when the number of observations  $n$  is small relative to the number of dimensions  $p$ . In order to solve this problem, we can work with mixtures of factor analyzers (MFA) as studied in section 3.4. After the application of this approach it is considered a reduction in the number of parameters through its factor-analytic representation of the component-covariance matrices. But it may not provide an appropriate reduction in the number of parameters, particularly when the number  $g$  of clusters (components) to be imposed on the data is not small. In section 3.6, we showed how we can reduce the number of parameters by using a factor-analytic representation of the component-covariance matrices with common factor loadings. The approach is called mixtures of common factor analyzers (MCFA). This sharing of the factor loadings enables the model to be used to cluster high-dimensional into many clusters and to provide low-dimensional plots of the clusters so obtained. The MFA approach does allow a more general representation of the component variances/covariances and places no restrictions on the component means. This model may not lead to a sufficiently large enough reduction in the number of parameters, particularly if  $g$  is not small. We support that MCFA provides a comparable approach that can be applied in situations where the dimension  $p$  and the number of clusters  $g$  can be quite large.

McLachlan et al. (2001), has underlined the significance of a mixture model-based approach to clustering as it yields a sound mathematical-based method. Although, in using this approach with mixtures of normal components that have nondiagonal covariance matrices, the number of observations to be clustered needs to be sufficiently large in number relative to their dimension in order to prevent singular estimates of the component-covariance matrices occurring during the estimation process. We have shown how we can handle this problem by adopting mixtures of factor analyzers to model the distribution of high dimensional vector of gene



expression data on a tissue. The aim of the application in data set Golub et al. (1999) was to underline the important role and usefulness of a mixture model-based approach to the clustering of microarray expression data. We have also discussed about how mixtures of factor analyzer models can identify various classes and subclasses among tissues on the basis of gene expression levels.

Baek and McLachlan (2008) claimed that in practice, we can use the Bayesian Information Criterion (BIC) of Schwartz to provide a guide to the choice of the number of factors  $q$  and the number of number of components  $g$  to be used. We can also use the likelihood ratio test statistic to choose  $q$ . For this reason we chose to perform the application of (i) two-component mixture factor analyzer with  $q=6$  factor, (ii) three-component mixture factor analyzer with  $q=6$  factor and (iii) two-component mixture factor analyzer with  $q=8$  factor, in chapter 4.



## References

- Baek J. and McLachlan G.J. (2008)** *Mixtures of Factor Analyzers with Common Factor Loadings for the Clustering and Visualisation of High Dimensional Data*, Department of Statistics, Chonnam National University, Gwangju 500-757, South Korea and Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia
- Bai J. and Ng S. (2008)**, *Large Dimensional Factor Analysis*, Foundations and Trends in Econometrics Vol. 3, No. 2 (2008) 89–163
- Bouveyron C. & Brunet - Saumard, C., (2014)** *Model-Based Clustering of High-Dimensional Data*, A review, Computational Statistics & Data Analysis, 71, 52-78.
- Broniatowski et al. (1983)** *Reconnaissance de densites par un algorithme d'apprentissage probabiliste*. In data analysis and informatics, vol. 3. Amsterdam: North-Holland, p. 359-374
- Bühlmann P., Sara van de Geer (2011)**, *Statistics for High Dimensional Data*, Springer
- Celeux G. and Diebolt J. (1985)**. *The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem*. Computational Statistics Quarterly 2, 73-82.
- Chang (1983)**, *on using principal components before separating a mixture of two multivariate normal distributions*. Applied Statistics, 32, 267-275
- Delyon, B., Lavielle, M. and Moulines, E. (1999)**. *Convergence of a stochastic approximation version of the EM algorithm*. The Annals of Statistics 27, 94-28.
- Dempster, A. P. et al. (1977)**. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society Series B 39, 1-38.
- Dudoit .S, Fridlyand,J. and Speed T.P. (2001)** *Comparison of discrimination methods for the classification of tumors using gene expression data*. J. Am. Stat. Assoc., to appear.
- Efron, B. and Hinkley, D. (1978)**. *Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information*. Biometrika 65, 457-487.
- Fraley, C. and Raftery, A. (1998)**. *How many clusters ?* The Computer Journal 41, 578-587.
- Gassiat, E. and Dacunha-Castelle, D. (1997)**. *Estimation of the number of components in a mixture*. Bernoulli 3, 279-299



**Ghahramani, Z. and Hinton, G. E. (1997)**, *The EM algorithm for factor analyzers*," Tech. Rep. CRG-TR-96-1, University Of Toronto, Toronto.

**Kass, E. and Raftery, A. (1995)**. *Bayes factors*. Journal of the American statistical Association 90, 773-795.

**Kogan J. (2007)** *Introduction to Clustering Large and High-dimensional Clustering*, Cambridge University Press

**Liu C. and Rubin D.B. (1998)**. *Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data*. Statistica Sinica 8, 729-747

**Louis T.A. (1982)**. *Finding the observed information matrix when using EM algorithm*, Journal of the Royal Statistical society B, 44, 226-233

**McLachlan G.J. , Bean R.W. and Peel D. (2002)** *A mixture model-based approach to the clustering of microarray expression data*, Bioinformatics, vol.18 no. 3 2002, pages 413-422

**McLachlan, G.J., and Brasford (1988)**. *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

**McLachlan G.J, Kim-Anh Do, Ambroise C. (2004)** *Analyzing Microarray Gene Expression Data*, Wiley Series in Probability and Statistics.

**McLachlan, G. J. and Krishnan, T. (1997)**, *The EM algorithm and extensions*, New York: John Wiley & Sons Inc.

**McLachlan G.J., Peel D. (2000)**, *Finite Mixture Models*, Wiley Series in Probability and Statistics.

**McNicholas P.D. & Murphy T.B.** *Parsimonious Gaussian Mixture Models*, Trinity College Dublin, Ireland

**Meilijson, I. (1989)**. *A fast improvement to the EM algorithm on its own terms*. Journal of the Royal Statistical society B, 51, 127-138

**Meng, X.-L. and VanDyk, D. (1997)**, *The EM algorithm - an old folk song sung to a new fast tune (with discussion)*," *Journal of the Royal Statistical Society, Series B*, 59, 511-567.

**Nouredinne El Karoui (2008)** *Spectrum Estimation For large dimensional covariance matrices using random matrix*, The Annals of Statistics 2008, Vol. 36, No. 6, 2757–2790



**Verleysen M. and François D. (2005)** *The Curse of Dimensionality in Data Mining and Time Series Prediction*

**Picard F. (2007)** *An introduction to mixture models*, Statistics For Systems Biology Group, Research Report No.7, March 2007

**Tipping, M. E. and Bishop, C. M. (1999a)**, *Mixtures of probabilistic principal component analysers*," *Neural Computation*, 11, 443-482.

**Wei, G. and Tanner, M. (1990)**. *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm*. *JASA* 82, 528-550.

**Wu, C. (1983)**. *On the convergence properties of the EM algorithm*. *The Annals of Statistics* 11, 95-103.

**Xiaochun Li, Ronghui Xu, (2009)** *High-Dimensional Data Analysis in Oncology*, Springer

