# Simultaneous Clustering with Mixtures of Factor Analysers

Author: Warwick O'Donnell
Supervisor: Dr. Maia Lesosky

November 2013

A DISSERTATION PRESENTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER OF SCIENCE DEGREE IN MEDICINE

University of Cape Town
Faculty of Health Science
Department of Medicine



**Abstract**

This work details the method of Simultaneous Model-based Clustering. It also presents an extension to this method by reformulating it as a model with a mixture of factor analysers. This allows for the technique, known as Simultaneous Model-Based Clustering with a Mixture of Factor Analysers, to be able to cluster high dimensional gene-expression data. A new table of allowable and non-allowable models is formulated, along with a parameter estimation scheme for one such allowable model. Several numerical procedures are tested and various datasets, both real and generated, are clustered. The results of clustering the Iris data find a 3 component VEV model to have the lowest misclassification rate with comparable BIC values to the best scoring model. The clustering of Genetic data was less successful, where the 2-component model could successfully uncover the healthy tissue, but partitioned the cancerous tissue in half.

## Acknowledgments

I would like to thank my supervisor, Dr Maia Lesosky, for her priceless guidance, advice, and for taking such an active role in this aspect of my life. Without her, this masters would simply never have been finished nor written. I could not have asked for a more helpful mentor.

Throughout my academic career, I have met many people who have helped me along my path. To start with, a thank you to Patrick Adams for all his help in getting me to masters. Throughout the years of undergrad, his patience and helpful nature has seen me through some academic struggles, and without whom my knowledge of MATLAB would be almost non-existent. This was very helpful throughout masters. I would also like to thank the people of The Shuttleworth Lap (TSL). That lab has a way of turning the most astute and proficient student in to a social bird whose academics are but a fleeting memory. Thank you all, and I hope to never see that place again. Last, but certainly not least, I would also like to thank Ryan Harmuth for all his help throughout the years. The sleepless nights of helping me with projects, proof-reading this masters and helping me through 2nd year physics. A very important component to my career, and a very caring friend.

I would also like to thank the cleaning people in lab K47 5.1. Only they, and my mother, know what a messy worker I can be. I thank the lab assistants for showing me how sleeping on the job is not as taboo as most of the hard workers of society claim it to be, and who's fruitful snores fill the air with a sweet musical note that rivals the complexity and majesty of Bach's most famous pieces. Thank you.

# Contents

# List of Tables



# List of Figures

# 1 Chapter 1: Introduction

## 1.1 Introduction

Thirty-something years in to the information age, the complexity and sheer volume of information necessitates sophisticated and automated data analysis techniques. The 'Information Age' is the phase in human history beginning somewhere around 1970 to the present day. This age is characterised by the sudden increase in available, consumed and manipulated information around the world. This can be seen in genetic data which poses many problems to the statistician due to it's high dimensional nature (McLachlan et al., 2001). It is possible that DNA can act as data storage devices (Church et al., 2012), however appropriate techniques are needed to retrieve this data. It is clear that in our age information is in abundance and, due to the high volume and dimensionality of this data, automatic techniques are necessary to handle this. Many such methods have been extensively studied; Regression analysis, factor analysis, principal component analysis, feature selection, discriminant analysis and cluster analysis to list a few.

Cluster analysis is an unsupervised method of data classification that allows the practitioner to partition data in to meaningful subgroups and uncover hidden subpopulations (eg. sex, race, species, ethnicity, etc) in the sample data. Cluster analysis is a form of unsupervised learning, where the only information that is known is the values of the features of unclassified data. That is to say that only the variables of each data point are known, with no assumed cluster number, group structure, or estimated parameters. The goal is to separate the data in to meaningful subpopulations in order to infer certain attributes and group structure of a population, or at least verify or discredit preconceived inferences. This form of data mining is contrasted with supervised learning, where a group of 'similar' samples are already classified and a rule is applied to the unclassified data that puts each observation in to the group that it has the highest probability of belonging to. Early approaches of cluster analysis involved heuristic approaches that involved arbitrary definitions of similarity and dissimilarity of subgroups or arbitrary distances from group centres (Ward, 1963; Macqueen, 1967). A parametric-based approach to clustering was defined in a probabilistic framework whereby each cluster is represented by a component distribution of a finite mixture model (Day, 1969). Mixture models are very useful in modeling a population sample that contains subpopulations. This allows statistical inferences to be made about the overall population by assuming properties attributed to each subpopulation. This is generally done by estimating the weights in the sum of components, and by estimating the parameters of each component probability density function that describes their respective sub populations. This technique of cluster analysis is called model-based clustering and has found application in the fields of genetics (Schork and Thiel 1996), (K. Y. Yeung et al. ,2001), medicine (McLachlan and Peel, 2000), tissue segmentation (Banfield and Raftery, 1993), diabetes diagnosis (Fraley and Raftery, 1998). The motivation for using this framework is that it allowed for a statistical inference and interpretation (Wolfe, 1970). As opposed to the former heuristic approaches, a question model-based clustering answered was "How many clusters?". The problem of cluster analysis reduced to one of model selection with measures such as the Bayesian Information Criteria (Schwartz, 1978) to assign a score to each model. Each model can be simplified by placing restrictions on various parameters. These model restriction produce, what is called, a parsimonious model. A parsimonious model, by the definition in the Dictionary of Common Concepts in Statistics, is said to be The simplest plausible model with the fewest possible number of variables. That is to say, that almost any statistical model whereby the number of parameters has been reduced, or restrictions have been placed on the parameters, is said to be a parsimonious model.

Model-based clustering with a mixture of any standard statistical distribution does not adequately model high dimensional data, such as that generated by modern genetics (Bouveyron and Brunet, 2012; Alladi

et al., 2008; Mclachlan et al., 2001). Problems related to data dimension and estimation occur when the number of observations (individuals) is much smaller than the number of variables (McLachlan et al, 2002). Secondly, the curse of dimensionality can be seen when estimating the covariance matrix of a model. The estimation techniques require an inversion of the covariance matrix, which can prove to be a computationally expensive task. One approach to dealing with high dimensional data is to utilise some form of data reduction technique. One such method is factor analysis which can be traced as far back as 1904 (Spearman, 1904). To model high dimensional data with a mixture model, a mixture of factor analysers is commonly used (McLachlan et al., 2001; Tipping and Bishop, 1998; McNicholas and Murphy, 2008). Many clustering algorithms simply fail to represent genetic data by an appropriate model due to over parameterisation (Day, 1969). Since then, many practitioners have developed methods to deal with this type of data. McLachlan et al (2001) used the mixture of factor analysers to model cancerous genes. Yeung et al. (2001) demonstrate how model-based methods are superior to older heuristic methods of modeling genetic data. Bailey and Elkan (1994) used model-based clustering to uncover various pieces of information in their analysis. It is necessary to implement rigorous mathematical techniques in order to deal with, and discover patterns in, genetic data. Model-based clustering with a mixture of factor analysers is a tool that has not only proved its worth in playing a vital role in creating a model to predict the genes that cause colon cancer (McLachlan et al., 2001), but shows much promise for future statistical analysis of genetic data.

Instead of modeling one sample of high dimensional gene expression data with a mixture of factor analysers, one may instead wish to model several samples. Each sample can be modeled separately and independently as usual, or one may wish to do so simultaneously by linking the models that represent each population sample. Finding a parametric link function between samples, in the supervised context, is not new in the field of population study (Van Franeker and Ter Brack, 1993; Biernacki *et al*, 2003). The idea of devising a mathematical link between population samples has recently been proposed in the unsupervised context of model-based clustering (Lourme and Biernacki, 2012). It is assumed from previous theory that the two populations are related. For example, the two populations may be considered to be "similar individuals described by the same features yet a difference exists between them due to time, location, ethnic group, etc" (Lourme and Biernacki, 2012). The proposed research will build on the work of (Lourme and Biernacki, 2012) where they introduce concepts for mixture models for linked data sets. It should be noted that the term 'simultaneous clustering' has two meanings in statistical literature. The first has been discussed above and the phrase 'simultaneous clustering' will refer to that method of clustering. The second common use for this phrase is defined as the clustering technique whereby instead of clustering the data matrix by either row or column separately, the rows and columns are clustered simultaneously. Other names given to the latter case are biclustering, two-way clustering, co-clustering or block clustering. This method of cluster analysis will not be studied in this dissertation.

The proposed advancement of the literature of simultaneous model-based clustering is to model several samples simultaneously with a mixture of factor analysers. This dissertation is therefore understood to be the defense of following claim:

> Simultaneous model-based clustering (Lourme and Biernacki, 2012) which allows one to model multiple similar population samples simultaneously by the use of a sufficiently simple inter-population link function, can be extended from a mixture of Gaussian distributions to a Gaussian mixture of factor analysers to account for overparameterised models.

This claim leads to a set of the following of sub-claims:

1. Simultaneous model-based clustering by a mixture of factor is desirable, useful and relevant.

2. There is a link function that links two similar populations that are modeled by a mixture of factor analysers, and the parameters of this mixture model can be estimated through the link function.

3. There is a suitable iterative procedure to estimate the parameters of the model.

The above claim was proposed by Lourme and Biernacki (2012) as a suitable extension to their method of simultaneous model-based clustering. The strategy for defending these claims will be as follows. First, sub-claim 1 will be addressed by showing that simultaneous model-based clustering by a mixture of factor analysers is a desirable and important technique in analysing high dimensional data. This is done by means of demonstrating how it could help the analysis of previous work in the field of genetics, and how historical work in the field mathematical statistics has lead up to this point. This is discussed in section 1.2. Chapter 2 covers the background literature and concepts required to understand simultaneous model-based clustering with a mixture of factor analysers. Broadly speaking, the chapter will first cover previous work on model-based clustering with some necessary parsimonious models (Banfield and Raftery, 1993), factor analysis as a data reduction technique with the associated parsimonious models (McNicholas and Murphy, 2008), and simultaneous model-based clustering (Lourme and Biernacki, 2012). Sub-claims 2 and 3 are discussed in chapter 3 where simultaneous model-based clustering with a mixture of factor analysers will be formulated to show that such a model is indeed possible and the parameters of which can be estimated with a suitable estimation procedure. Finally this model will be tested in Chapter 4 to verify it‘s usefulness. Chapter 5 discusses and interprets the results, and gives a list of the strengths and weaknesses of the model. It also discusses possible future work.

## 1.2 Literature Review

### 1.2.1 Review of the Applications of Model-based Clustering of Biological Data

This section introduces the reader to some of the background literature where model-based clustering was performed on biological data. The analysis of genetic data is notoriously difficult because of the number of variables relative to the number of samples (McLachlan et al, 2001) which makes for very noisy data. The aim of this section is to demonstrate that simultaneous model-based clustering by a mixture of factor analysers can potentially have application in the health sciences with specific application to overparameterised data structures such as gene expression data.

Bailey and Elkan (1994) used model-based clustering to discover motifs in bipolymers. They used a two-component mixture model, the parameters of which were estimated using the EM algorithm as outlined by Aitkin and Rubin (1985) whereby the mixing proportions and the parameters of the underlying distribution are to be estimated iteratively. The one component is designed to explain a set of similar sub sequences of fixed width (“the motif”). The second component is constructed to describe all the positions in the sequences. To solve this they use what is called the MM algorithm. This algorithm uses the EM iterative procedure, however differs slightly to that described by (Lawrence and Reilly, 1990) in that it relaxes particular assumptions relating to the number of occurrences of the motif pertaining to the data set. Some more useful reviews can be found in Fraley and Raftery (2002), Fraley and Raftery (1998) and Melnykov and Maitra (2010).

Yeung et al. (2001) attest to the usefulness of clustering high dimensional micro array data in order to interpret and exploit it. They demonstrate how model-based approaches to clustering are favourable over the older heuristic approaches such as k-means and graph theoretic approaches. They note that with the usual heuristic approach the number of clusters is difficult to uncover and it is difficult to define what a “good” clustering algorithm is. With model-based approaches the clustering algorithm reduces to a model selection

process and choosing a suitable clustering method. This, they comment, is a great advantage over heuristic approaches where there is no general way of determining the number of clusters or which heuristic method is better. From their analysis on two sets of gene expression data sets, and three synthetic data sets they conclude that the VVV model, for the synthetic data modeled by a mixture of normal distributions, produced the highest quality clusters and the BIC chose the right model and number of clusters. For the randomly sampled synthetic data, the diagonal covariance structure was the superior model. Their conclusions were that model-based clustering was the superior method when modeling data similar to gene expression data.

McLachlan, Bean and Peel (2001) introduce the software EMMIX-GENE that has gene selection process step before clustering the gene expression data. In their paper, they introduce a method that selects the most important genes involved in obtaining the ones that are most predictive in determining cancerous cells. This method, in turn, reduces the dimensions of the data which eliminates a lot of noise form the data set. In their analysis they use the colon cancer data set from Alon et al. (1999) which originally had over 6500 gene expressions from 40 samples of normal tissues and 22 samples of tumour tissue. Alon et al (1999) focuses on the 2000 genes with the highest minimal intensity of the sample. The data matrix is $2000 \times 62$. McLachlan, Bean and Peel (2001) introduce the likelihood ratio statistic as a screening process of the relevant genes. In this they apply the statistic $-2ln(\lambda)$ to test a 1 component versus 2 component t-distributed mixture model. Here $\lambda = L(g = 2) - L(g = 1)$ and where $L(g = i)$ is the likelihood function with number of components equal to $i$. This statistic is tested for each gene across each tissue. If the difference in the likelihoods is significant then it is taken that the specific gene which is tested is relevant in classifying cancerous tissue. With this screening process they uncover, from the 2000 genes originally considered, 446 genes were considered relevant. A 2-component mixture model was fitted to the data with the number of factors ranging from 2 to 8. They remark that there was little difference in each model but that a 6 factor model was favourable.

Pournara and Wernisch (2007) used the techniques of implementing a factor analytic model, as outline by Ghahramani and Hinton (1996), to uncover unobserved variables in the structure of gene regulatory networks. These unobserved variables are known as Transcription Factors (TF), and are uncovered by a two-layer network, the first of which is the unobservable TF and the second are the observed gene expression variables. Pournara and Wernisch (2007) analysed 5 different factor analytic algorithms, Bayesian and classical, and compare the results.

Lourme and Biernacki (2012) tested their model of simultaneous model-based clustering on three populations of Cory‘s Shearwater bird species (Thibaults et al, 1997). There were measurements conducted of three different species (H=3) of 5 different variables. The goal is to successfully uncover a 2-component cluster in all three samples, where each cluster of each sample indicates males or females. Each cluster has the same meaning in each sample. The study showed a comparison between the independent and simultaneous clustering method and showed that the best scoring simultaneous method out-performed the best scoring independent clustering method for G=2 groups. It was also confirmed that when the clustering number is not known the simultaneous method outperformed the independent method for G=1, 3 and 4.

It is clear that simultaneous model-based clustering produces consistent results for low dimensional data. To the author‘s knowledge no tests have been done using simultaneous model-based clustering with a mixture of factor analysers. This could be useful for analysing high dimensional gene expression data.

### 1.2.2 Review of the Mathematical Literature

The historical problems of modeling high dimensional data, such as those of genetic data discussed in section 1.2.1, will also be reviewed with specific attention on factor analysis. This section also serves to demonstrate that the past and present literature has lead up to the natural extension of simultaneous model-based clustering by a mixture of factor analysers.

One of the first pieces of work on model-based clustering is done by Day (1969). In his paper he discusses a two component mixture model with equal component covariance matrices. He then compares several estimation techniques; moment estimators, minimum $\chi^2$ estimation, Bayes estimators and the popular maximum likelihood estimation. A generalised finite mixture model is then proposed where he tests the implications, on each estimation technique: of having unequal component covariance matrices and more than two components. Day concludes that finite mixture models could be superior in clustering data to previous clustering techniques. He further concludes that maximum likelihood estimation is the superior parameter estimation technique for multivariate data, but this technique becomes computationally exhaustive, and sometimes impossible, for dimensions $k > 10$. This problem, refereed to as "the curse of dimensionality" by Bellman, gets explored years later. Day's implementation of the method of maximum likelihood will be demonstrated, and his results leading to his conclusions will be given.

Day formulates the maximum likelihood equation of two normal distributions with equal covariance matrices as follows; Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be means of component 1 and 2 respectively, and $\boldsymbol{\Sigma}$ the common covariance matrix and $p$ the constant of proportionality. Then the likelihood equation is given by

$$l(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, p, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2}nk}|\boldsymbol{\Sigma}|^{-\frac{1}{2}n}\prod_{i=1}^{n}[pe^{-\frac{1}{2}(\mathbf{x}_i-\mu_1)\Sigma^{-1}(\mathbf{x}_i-\mu_1)^T} + (p-1)e^{-\frac{1}{2}(\mathbf{x}_i-\mu_2)\Sigma^{-1}(\mathbf{x}_i-\mu_2)^T}]$$

After maximising the log likelihood equation with respect to $\boldsymbol{\mu}_1$,$\boldsymbol{\mu}_2$, $p$ and $\boldsymbol{\Sigma}$ the following update equations are given

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_1 &= \frac{\Sigma_j \mathbf{x}_j \hat{P}(1|\mathbf{x}_j)}{\Sigma_j \hat{P}(1|\mathbf{x}_j)} \\
\hat{\boldsymbol{\mu}}_2 &= \frac{\Sigma_j \mathbf{x}_j \hat{P}(2|\mathbf{x}_j)}{\Sigma_j \hat{P}(2|\mathbf{x}_j)} \\
\hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\Sigma_j[(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_1)^T(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_1)\hat{P}(1|\mathbf{x}_j) + (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)^T(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_2)\hat{P}(2|\mathbf{x}_j)]
\end{aligned}$$

where $\hat{P}(k|\mathbf{x}_j)$ is the probability of the $j$th observation belonging to component $k$ and

$$\begin{aligned}
\hat{P}(1|\mathbf{x}_j) &= \frac{\hat{p}e_{1j}}{\hat{p}e_{1j} + (\hat{p}-1)e_{2j}} \\
\hat{P}(2|\mathbf{x}_j) &= 1 - \hat{P}(1|\mathbf{x}_j)
\end{aligned}$$

where $e_{ji} = e^{-\frac{1}{2}(\mathbf{x}_i-\mu_j)\Sigma^{-1}(\mathbf{x}_i-\mu_j)^T}$.

The results shown by Day (1969) demonstrate that the method of moment estimation yielded poor results when estimation the parameters of the multivariate finite mixture model was attempted, compared to the method of maximum likelihood as shown above. This paper gave justification as to why the method of maximum likelihood estimation is the preferred method of parameter estimation of finite mixture models.

Day further discussed the cases of more than two components and concluded that the method of maximum likelihood is still a viable option for estimation. However, he found that maximum likelihood procedure

breaks down when one deals with unequal covariance matrices and especially when the number of dimensions is greater than 10, although his analysis of this case was no as thorough as subsequent papers. Day found that the likelihood becomes infinite for data with too large a dimension. In conclusion, Day found that finite mixture models (with the above equations as parameter estimations) is a superior way of clustering when compared to previous methods. He gave three reasons. Firstly, clustering can be performed irrespective of the size of the sample. Secondly, A general covariance matrix may be assumed, which takes in to account any linear relationships that may exist between variables. Thirdly, the distribution of the proportion variable $\hat{p}$ and the generalized distance between clusters variable $\hat{\Delta} = [(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)]^{\frac{1}{2}}$ can be estimated and used to test significance of clusters. The disadvantages found were that as the number of dimensions increases, the computation of the maximum likelihood becomes increasingly prohibitive. The second disadvantage found was the the clusters are not so clear cut because each cluster was defined by a probability distribution. He comments, however, that this could turn out to be an advantage.

J. H. Wolfe (1970) officially reformulated cluster analysis by assuming that each cluster belongs to a component distribution of a finite mixture model. In doing this he removed the arbitrary definitions of 'similarity', that was the foundation for cluster analysis at the time, and introduced more of a statistical interpretation. In his paper, J. H Wolfe (1970), he presented the general theory of finite mixture models and demonstrated the general procedure for parameter estimation of the weighting parameter and the parameters associated with the component distribution. He demonstrated how finite mixture models apply to cluster analysis and estimates parameters by maximum likelihood. His choice for this estimation procedure could very well be due to the findings of Day (1969). Another difference to Day (1969) is that in J. H. Wolfe (1970) a simple formal procedure is described that determines the number of clusters. Here, the number of clusters, $\mathbf{r}$, was tested against the alternative hypothesis with number of clusters,$\mathbf{r}'$, by evaluating the likelihood ratio $\chi^2 = -2ln(\frac{L_{\mathbf{r}}}{L_{\mathbf{r}'}})$ with degrees of freedom equal to the difference in the number of parameters estimated. This likelihood ratio test may also be used when testing the distribution of components against each other (J. H. Wolfe, 1970).

The estimation of the parameters of a normal mixture model were presented in theorem 3 in J. H. Wolfe (1970) for unequal covariance matrices, given by

$$\begin{aligned}
\hat{\pi}_s &= \frac{1}{N}\sum_{k=1}^{N}\hat{P}(s|\mathbf{x}_k) \\
\hat{\mu}_{si} &= \frac{1}{N\hat{\lambda}_s}\sum_{k=1}^{N}\hat{P}(s|\mathbf{x}_k)\mathbf{X}_{ik} \\
\hat{\sigma}^s_{ij} &= \frac{1}{N\hat{\lambda}_s}\sum_{k=1}^{N}\hat{P}(s|\mathbf{x}_k)(\mathbf{X}_{ik} - \hat{\mu}_{si})(\mathbf{X}_{jk} - \hat{\mu}_{si})
\end{aligned}$$

where $\hat{\pi}_s$ is the mixing proportion estimate of cluster $s$, $\hat{P}(s|\mathbf{x}_k)$ is known as the "probability of membership" of cluster $s$ and is given by $\frac{\pi_s\alpha_s(\mathbf{x},\Theta)}{f(\mathbf{x})}$ where $\alpha_s(\mathbf{x},\Theta)$ is the component distribution with parameters $\Theta$ and $f(\mathbf{x}) = \sum_{s=1}^{\mathbf{r}}\pi_s\alpha_s(\mathbf{x},\Theta)$. The mean of component $s$ is given by $\hat{\mu}_{si}$, and component covariance matrix is $\hat{\sigma}^s_{ij}$.

J. H. Wolfe (1970) tested his theorem on the Iris data set publish by Fisher (1936). In his analysis he found that a 3 component mixture model outperformed the two component and one component with a misclassification of 3 of the flowers. However, it was unclear if the three component or the four component model was superior. He concluded that his results for the likelihood test show that his method can be improved upon better approximating the likelihood ratio expression.

Later Scott and Symons (1971) used the idea of assigning each cluster to one of a finite number of probability distributions. In their paper they tackled the problem of estimating the parameters of the distribution without having any prior knowledge of this distribution underlying the data. This was novel at its time since preceding classification techniques assumed the knowledge of the underlying component distributions to be known, or at least an abundance of information about the data was given (Scott and Symons, 1971). In their paper, a component labeling was introduced where the component label $\boldsymbol{\gamma}$ (relabelled as $\boldsymbol{z}$ in this dissertation) is a parameter of $n$ components such that the $i$th component indicates which cluster, or distribution, the $i$th observation belongs. Their analysis was the same as that of Day (1969) except for the additional assumption that $\boldsymbol{z}$ is an unobservable random variable, the components of which are the outcomes of $n$ multinomial trials. Scott and Symons (1971) defined the likelihood differently to Day (1969). They defined a general likelihood with respect to previously clustered data as well as data with no prior knowledge available. This provides a likelihood equation that allows for supervised, as well as unsupervised, classification. Their likelihood is defined as follows; Let $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ be a set of $n$ $p$-variate observation. Each observation is assumed independent and may arise from any one of $G$ multivariate distributions with means $\boldsymbol{\mu}_1, ...., \boldsymbol{\mu}_G$ and covariances $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_G$. Let $\mathbf{x}_{g1}, ..., \mathbf{x}_{gm_g}$ be a set of independent observations belonging to on of the $G$ distributions. The classification parameter $\boldsymbol{z} = (z_1, ..., z_n)$ is such that $z_i = g$ if $\mathbf{y}_i$ comes from cluster $g$. The parameter set $\boldsymbol{\theta} = \{\boldsymbol{z}, \boldsymbol{\mu}_1, ...., \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_G\}$ is to be estimated from the likelihood equation given by

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{G}[\sum_{i=1}^{m_g}(\mathbf{x}_{gi} - \boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_{gi} - \boldsymbol{\mu}_g) + \sum_{C_i}(\mathbf{y}_i - \boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_g) + (m_g + n_g)ln|\boldsymbol{\Sigma}_g|]$$

where $C_g$ is the set of $\mathbf{y}_i$'s assigned to cluster $g$ by $\boldsymbol{z}$, $n_g$ are the number of observations in cluster $g$.

The classification problem, supervised or unsupervised, is to estimate $\boldsymbol{z}$ and therefore cluster the $C_i$'s. The supervised classification occurs when the covariances and means are known, or a sufficient number of samples are given for each sub population. In the case of unsupervised classification, no prior knowledge is assumed, nor any previous samples given. Scott and Symons (1971) addressed the problem posed in Day (1969) where Day found that the maximum likelihood procedure breaks down by having an infinite likelihood when the number of variables is too large. Scott and Symons (1971) added the criterion that a minimum of $p + 1$ observations must be assigned to each cluster to avoid the degenerate case. This was the first time this criterion, or anything like it, was implemented. Scott and Symons (1971) reformulated the test of the number of clusters by testing the null hypothesis $H_0 : z_1 = z_2 = ... = z_n$ against the hypothesis that not all the $z_i$'s are equal. In this way they can test the number of clusters. A comparison between the method of Scott and Symons (1971) and Day (1969) is given in F. Marriott (1975).

In later work Symons (1981) introduced new criteria for the case in model-based clustering of a mixture of multivariate normal distributions. These criteria are imposed on the likelihood equation in the cases of equal and unequal component covariance matrices. The likelihood is given by

$$l(\mathbf{Y}, \boldsymbol{\theta}, \mathbf{z}) = \prod_{g=1}^{G} \pi_g^{n_g}|\boldsymbol{\Sigma}|^{-\frac{1}{2}n_g}e^{-\frac{1}{2}(\sum_{g=1}^{G}\sum_{C_g}(\mathbf{y}_i - \boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_g))}$$

where $C_g$ indicates the cluster to which $\mathbf{y}_i$ belongs and $z_i$ equals $g$ if $\mathbf{y}_i$ belongs to the $g$th component.

The maximum likelihood estimates are given by

$$\begin{aligned}
\pi_g &= \tfrac{n_g}{n} \\
\hat{\boldsymbol{\mu}}_g &= \tfrac{1}{n_g}\textstyle\sum_{C_g}\mathbf{y}_i \\
\hat{\boldsymbol{\Sigma}}_g &= \tfrac{1}{n_g}\boldsymbol{W}_g \qquad = \dfrac{1}{n_g}\displaystyle\sum_{C_g}(\mathbf{y}_i-\hat{\boldsymbol{\mu}}_g)(\mathbf{y}_i-\hat{\boldsymbol{\mu}}_g)^T
\end{aligned}$$

for $g = 1, ..., G$. The first criteria is for the case of equal component covariance matrices $\Sigma_g = \Sigma$ with $\Sigma$ unknown. In this case the estimation equation changes to

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{W} = \frac{1}{n}\sum_{g=1}^{G}\boldsymbol{W}_g$$

The optimal allocation of the observation points given by $\hat{\mathbf{z}}$ maximises the criterion

$$nln|\hat{\boldsymbol{W}}| - 2\sum_{g=1}^{G} n_g ln(n_g)$$

For the case of unequal component covariance matrices $\Sigma_g$ this criterion which must be maximised is

$$\sum_{g=1}^{G} n_g ln|\hat{\boldsymbol{W}}_g| - 2\sum_{g=1}^{G} n_g ln(n_g)$$

When dealing with clusters of equal size, the above criteria work very well, and suggests that the above criteria works well for different shaped clusters (Symons, 1981).

The method was used on a data set consisting of 145 non-obese patients to uncover a relationship between chemical diabetes and overt diabetes. Their technique uncovered three different-shaped clusters each representing the patients with chemical diabetes, overt diabetes and normal subjects. The conclusion in Symons (1981) is that it is unclear if the above criteria produced a superior estimation to the classification of this particular data set, but that the technique presented above produces adequate clusters, which was verified by previous works that clustered similar data.

In more recent work Banfield and Raftery (1993) devised a reparameterisation of the component covariance matrices in the context of model-based clustering that allows the orientation, size and shape of the clusters to be the same or different. Their proposed reparameterisation helps when trying to estimate highly parameterised models. The reparameterisation is given by

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{F}_g \mathbf{A}_g \mathbf{F}_g^T$$

where $\lambda_g$ is the first eigenvalue of $\boldsymbol{\Sigma}_g$, $\mathbf{A}_g = diag\{\alpha_{1g}, ..., \alpha_{dg}\}$ with $1 = \alpha_{1g} \geq ... \geq \alpha_{dg} \geq 0$ and $\mathbf{F}_g$ the matrix of eigenvectors. The maximum likelihood estimation procedures at the time only accounted for equal component covariance matrices, or the unparsimonious model of arbitrary covariance matrices. The reparameterisation above of the covariance matrix allowed the user more freedom with regard to the similarity of the shape, size and orientation of each cluster. The $\lambda_g$ controlled the space that the cluster takes in $p-$dimensional space, $\mathbf{F}_g$ describes the orientation of the clusters and $\mathbf{A}_g$ describes the shape. The mclust package in R uses this eigen-decomposition to control the features of the clusters. The myriad of models described in Banfield and Raftery (1993) have formed a basis for parsimonious mixture models in future work and is used in the modeling presented in this dissertation.

Some decades later Lourme and Biernacki (2012) presented a method to cluster several samples of data, using Gaussian mixture models, simultaneously. They did this by finding the link function that transforms, in distribution, each component of one mixture model in to their respective components of a model in a different population. Their work is presented in more detail in section 2.4 and 2.5.4. Their idea was novel for two reasons. Firstly, finding a link between two sample has only been done in a supervised context. Lourme and Biernacki (2012) propose a solution where they find a link function between two samples, in the case for cluster analysis, and successfully cluster the second sample without estimating the covariances and means of the model of the second sample. This is done by estimating the parameters of the link function, and then simply calculating the covariances and means of the second sample using the parameters of the link function and reference population. The second reason it was novel was because it was the first of its kind to cluster data across populations. Previous works at finding links was done between sample of the same population, whereas Lourme and Biernacki (2012) showed that one can find a link between two different populations provided the individuals of each populations are similar and can be described by the same variables. Note that Biernacki et al. (2003) also found a link between two different populations but this was done in a supervised context. Their results showed that the model of simultaneous clustering outperformed some usual independent clustering techniques. Below gives the results of clustering several samples of Cory‘s Shearwaters.

Table 1: *Best BIC values obtained from the simultaneous versus independent clustering*

| Cluster number | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Simultaneous | 4047.8 | **4047.0** | 4051.0 | 4055.7 |
| Independent | 4102.6 | 4139.8 | 4137.7 | 4159.6 |

They also found that when the core assumption, that each population are to have the exact same descriptors, is relaxed the model still performed feasibly.

One year later Biernacki and Lourme (2011) presented the method for simultaneous clustering of a mixture of t-distributions. Their solution was the same as presented in Lourme and Biernacki (2012), except for an additional constraint that the degrees of freedom are equal across populations. That is $\nu_{k=\ldots\ldots\ldots=}^{1}\nu_k^H$ for $k \in \{1, \ldots, K\}$ of $K$ groups and $\nu_k^h$ is the degrees of freedom of component $k$ and population $h$. Their method was tested on the status of companies and whether or not they are bankrupt. They test four ratios of companies properties and clustered the data where the desire is find a two component mixture model one indicating healthy companies and the other indicating bankrupt companies.

Table 2: *Best ICL values obtained from the simultaneous versus independent clustering*

| Cluster Number | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Simultaneous | −1169.7 | −1191.3 | **−1202.0** | −1183.4 | −1131.3 |
| Independent | −1154.6 | **−1163.6** | −1072.1 | −1127.7 | −1098.3 |

The algorithm found a three component model. The associated confusion table shows the misclassifications.

Table 3: *Confusion table of simultaneous clustering*

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| Healthy | 3 | 94 | 360 |
| Bankrupt | 56 | 10 | 366 |

Cluster one and cluster two clearly represent bankrupt and healthy companies respectively. Cluster three

represented the model uncertainties about bankruptcy status.

The curse of dimensionality has been tackled as far back as 1904 (Spearman, 1904). In his paper he discussed a common factor underlying the grades of school pupils. He labels this factor "general intelligence". He believed that his data points of each pupil's school subject marks were all highly correlated with with this factor. Since then much work has been done on devising a mathematical approach to dimension reduction by clustering the data in a lower subspace. As Bouveyron and Brunet (2012) point out, dimension reduction techniques are superior to the reparameterisations such as banfield and Raftery (1993) when dealing with the case for unsupervised classification. Furthermore, the approach of dimension reduction has been given much credit by the works of Huber (1985) where he discovered certain useful properties of high dimensional spaces. In particular he found that high dimensional spaces are mostly empty. The experiment is as follows; Assume a p-variate random vector $\mathbf{Y}$ with uniform density over some hypersphere of radius 1 is given. The probability that some realisation of this random vector $y_i$ is between this sphere and the hypersphere of the same dimension and of radius 0.9 is given by

$$P(y_i \in S_{0.9}(p)) = 1 - 0.9^p$$

As an example, the probability that a 30 dimensional data point belongs to the above shell is $1 - 0.9^{30} \approx 0.9576$, which means that most of the data points live in a $p-1$ space and the rest of the space is almost empty. This shows that clustering data in a lower dimension could easily lead to favourable results without the loss of too much information.S

There are many dimension reduction approaches, but the one considered in this paper is factor analysis (Spearman, 1904). Probabilistic principal component analysis is a specific form of factor analysis. Spearman defines principal component analysis as the "linear projection that minimizes the average projected cost". Later Hotelling (1933) redefined PCA to be the reduction is the dimension of the data while preserving as much of the variation as possible. The principal vectors turn out to be the eigenvectors associated with the largest eigenvalues of the covariance matrix.

In more recent work, Ghahramani and Hinton (1997) introduced the mixture of factor analysers. This model allows a factor model to be locally applied to different regions of the data space whereas the usual factor model assumes common factors amongst all data points. Their work introduced the MFA (Mixture of Factor Analysers) model in the context of unsupervised classification where they combined dimension reduction (factor analysis) and model-based Gaussian clustering. There are two major benefits of local dimension reduction versus the idea of clustering and then reducing dimension separately. Firstly, it may be that different features may be correlated within different clusters and therefore each cluster may have different factors. Secondly, they propose that with this method different clusters may appear more separated depending on the local metric. The algorithm of Ghahramani and Hinton (1997) was an adaptation of Hinton et al. (1996) where they have a two step algorithm; First cluster the data (outer loop), then apply each individual factor model (inner loop). Ghahramani and Hinton (1997) present the EM algorithm which removes the need for an inner and outer loop and in turn reduces the number of heuristic parameters. The single factor analystic model they proposed is as follows; assume a p-dimensional real valued data vector $\mathbf{x}$ is modeled using a real valued factor vector $\mathbf{z}$ of dimension $k$ where $k < p$. The model is then given by

$$\mathbf{x} = \Lambda\mathbf{z} + \mathbf{u}$$

where $\Lambda$ is the factor loading matrix (referred to, henceforth, as $\mathbf{B}$). The factors $\mathbf{Z} \sim N(0, I)$ (referred to

henceforth as $\mathbf{U}$), and p-dimensional random variable $\mathbf{u} \sim N(0, \psi)$ (henceforth referred to as $\epsilon$) where $\psi$ is a diagonal matrix - a key assumption in factor analysis that the variables are independent given the factors. Given this model we have $\mathbf{X} \sim N(0, \mathbf{BB}^T + \psi)$. The goal, therefore, is to find $\mathbf{B}$ and $\psi$ that best describes the data. The EM algorithm is presented in sections 2.5.2 .
For the case of mixture of factor analysers, Ghahramani and Hinton present the following generative model

$$Y_{|Z=k} = \mathbf{B}_k \mathbf{U} + \epsilon$$

where the factors are all assumed to be distributed normally with mean zero and covariance as the identity matrix and $\epsilon \sim N(0, \psi)$. In their analysis they set the individual specific variances to be equal across components, $\psi_i = \psi$.

Tipping and Bishop (1998) introduced PPCA. They noted one of the limitations of PCA to be that it offers no probabilistic model, only a dimension reduction. This model is similar to the factor analytic model presented in Ghahramani and Hinton (1997) except they make the restriction on the covariance matrix $\Sigma$ to be

$$\Sigma = \mathbf{BB}^T + \sigma^2 I$$

In this way the PPCA model can be seen as special case of the factor analytic model.

The work of Ghahramani and Hinton (1997) was later generalised by McLachlan et al. (2002), although this model was also considered in McLachlan and Peel (2000), where they considered the case of unequal variance of the noise where we have the conditional distribution of the noise term to be $\epsilon|Z \sim N(0, \psi_k)$ where $\psi_k$ is the diagonal covariance matrix of cluster k. The model they propose is given as follows

$$\mathbf{Y}_j = \boldsymbol{\mu}_i + \mathbf{B}_i \mathbf{U}_{ij} + e_{ij}$$

where $\mathbf{Y}_j$ is a data vector, $\boldsymbol{\mu}_i$ is the centre of component $i$, $\mathbf{B}_i$ is the factor loading matrix of component $i$ and $\mathbf{U}_{ij}$ the factor scores and where the noise term is normally distributed as $e|Z \sim N(0, \psi_k)$ where $\psi_k$ is the diagonal covariance matrix of cluster $k$. McLachlan et al. (2002) demonstrate the reduction in parameters offered by the factor analytic model. They note that the restriction that $\mathbf{B}^T \psi^{-1} \mathbf{B}$ needs to be diagonal implies that the factor analytic model have $pq + p - \frac{1}{2}q(q-1)$ free parameters to estimate. . While Ghahramani and Hinton (1997) devise their own EM algorithm procedure to estimate parameters, McLachlan et al. (2002) make use of the alternating expectation conditional maximisation procedure and use the result that $(\mathbf{BB}^T + \psi)^{-1} = \psi^{-1} - \psi^{-1}\mathbf{B}(I_q + \mathbf{B}^T\psi^{-1}\mathbf{B})^{-1}\mathbf{B}^T\psi^{-1}$ and as shown in section 2.5.3.

McNicholas and Murphy (2008) summarised the work of parsimonious Gaussian mixture models by including the MFA model of Ghahramani and Hinton (1997), McLachlan et al. (2002), Tipping and Bishop (1999) and some of their own models. In their paper, they generalised parsimonious Gaussian mixture models and give a table of possible parsimonious models. They give a full range of constraints across groups for the factor loading matrix $\mathbf{B}_i$, noise term $\psi_i$ and whether or not $\psi_i = \sigma_i \mathbf{I}$ . The PGMMs of McNicholas and Murphy (2008) are discussed further in section 2.3.4. Their parameter estimation involved the use of the alternating expectation conditional maximization algorithm as can be seen in their paper. Using their 8 models on two different data sets, McNicholas and Murphy (2008) find that with their PGMM the number of parameters grows linearly with a growth in data dimension, while the the parameters of the Gaussian mixture model, offered in mclust, grows quadratically in the case of a non-diagonal covariance matrix. This shows that the PGMM has potentially more flexibility when clustering high dimensional data. Through their experiments they further add that the PGMMs offer a more feasible solution when the variables are

highly correlated, and these models can outperform previous attempts at data classification.

Later Baek et al. (2009) devised the Mixture of common factor analysers model. This model offered a more parsimonious type of modeling where they restrict the mean and the covariance matrices to allow for an even larger reduction in estimated parameters. As usual, the factor analytic model is such that the covariance matrix $\Sigma_i = \mathbf{B}_i\mathbf{B}_i^T + \psi_i$ where $\mathbf{B}_i$ is the factor loading matrix and $\psi_i$ the error covariance matrix of component $i$. The proposed restriction is given by

$$\boldsymbol{\mu}_i = \mathbf{A}\boldsymbol{\xi}_i$$

$$\Sigma_i = \mathbf{A}\boldsymbol{\Omega}_i\mathbf{A} + \psi$$

for $p$-dimensional data and $q$ factors, and where $\mathbf{A}$ is a $p \times q$ matrix, $\boldsymbol{\xi}_i$ is a $q$-dimensional vector $\boldsymbol{\Omega}_i$ is a $q \times p$ positive definite symmetric matrix and $\psi$ is $p \times p$ diagonal. In this case the data vectors are modeled as

$$\mathbf{Y}_j = \mathbf{A}\mathbf{U}_{ij}^* + \epsilon_{ij}$$

with probability $\pi_i$ for $i = 1, ..., g$ and $j = 1, ..., n$. In this case the unobservable factors $\mathbf{U}_{ij}^*$ are distributed independently as $N(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i)$, independently of $\epsilon_{ij}$ which is distributed as $N(\mathbf{0}, \psi)$. The matrix $\mathbf{A}$ is the $p \times q$ factor loading matrix which are chosen to satisfy the relation $\mathbf{A}\mathbf{A}^T = \mathbf{I}_q$. The difference between the MCFA model and a the MFA model is that in the MCFA, the data points $\mathbf{Y}_j$ are modeled directly whereas in the MFA model it is the centred data $\mathbf{Y}_j - \boldsymbol{\mu}_i$ that is modeled. Secondly, the MCFA model is a specific type of MFA model where

$$\begin{aligned}
\boldsymbol{\mu}_i &= \mathbf{A}\boldsymbol{\xi}_i \\
\mathbf{B}_i &= \mathbf{A}\mathbf{K}_i \\
\mathbf{U}_{ij}^* &= \mathbf{K}_i^{-1}(\mathbf{U}_{ij} - \boldsymbol{\xi}_i) \\
\psi_i &= \psi
\end{aligned}$$

where $\mathbf{U}_{ij}$ is the factor scores of the MFA model and where $\mathbf{U}_{ij}^* \sim N(0, \mathbf{I}_q)$. $\mathbf{K}_i$ is chosen such that $\mathbf{K}_i^{-1}\boldsymbol{\Omega}_i(\mathbf{K}_i^{-1})^T = \mathbf{I}_q$ which ensure that $\mathbf{U}_{ij}^* \sim N(0, \mathbf{I}_q)$. This model can essentially be seen as the MFA model proposed by Ghahramani and Hinton (1997) but with common factor loadings $\mathbf{A}$ and transformation matrix $\mathbf{K}_i$ to ensure the correct distribution of $\mathbf{U}_{ij}^*$. The MFA model is a better model than the MCFA provided the data reduction is adequate. That is if we reduce the data from $p$ dimensions to $q$ dimensions where the the number of parameters in the subspace allows for a manageable model. The MCFA is a more feasible if the reduced subspace of the MFA model still has too many parameters.

It is clear that the mathematical literature has lead up to the point where one can implement simultaneous model-based clustering by a mixture of factor analysers which can be utilised in the field of genetics. To the authors knowledge, there has been no work done on this. This dissertation presents the case for the simultaneous model-based clustering with a mixture of factor analysers and gives the estimation procedure for this model. Several parsimonious models will be presented and some of them will be tested.

# 2 Chapter 2: The Mathematical Background

This chapter covers the mathematical preliminaries required to understand simultaneous model-based clustering with a mixture of factor analysers. Section 2.1 reviews finite mixture models, focusing on the Gaussian mixture model. Gaussian models are widely used because of many simplifying properties and extensive parameter estimation methods available. Parameter estimation will be implemented by the EM algorithm and some extensions there of. The extension of GMM known as parsimonious GMM (PGMM) will be covered including detailed discussion of the covariance matrix decomposition. Finally a short account of the indentifiability problem will be given, which is an important problem when dealing with linked data sets.

Section 2.2 focuses on data reduction techniques, mainly factor analysis. Factor analysis is a key concept in dimension reduction methods and important in the following work. . It‘s role as a parsimonious mixture model in model-based clustering and the interpretation of the latent variables known as factors will be discussed. In performing factor analysis a reduction in the number of parameters is achieved by placing restrictions on the covariance matrix. The EM algorithm procedure will be given for this problem.

Section 2.3 covers simultaneous model based clustering (SMBC) (Lourme and Biernacki, 2012) with discussion of the population link parameters. This gives the solution to the transformation equation that is needed to transform the mixture model of the reference sample (the sample that has been fully classified) in to that of another different, but similar, sample. The idea is to classify a sample by finding the link between it and an already classified sample set. This link is defined by it‘s unique transformation parameters. Restrictions on the model will be proposed and the parameter estimation procedure given.

Section 2.4 will introduce the definition and application of overlap of clusters and section 2.5 is an overview of the EM algorithm as it pertains to the mixture of Gaussians model, factor analytic model and simultaneous clustering model.

## 2.1 Finite Mixture Models

A finite mixture model is a probability density function that is a finite weighted sum of other probability density functions (pdf). Individual pdfs are known as the ‘component‘ densities for the mixture model. The aim of model-based clustering is to fit a mixture model to a population and identify each component with a cluster, where each cluster represents a sub population. The number of clusters is determined from the data, usually using the BIC score.

The pdf of a G-component mixture model is defined as

$$f(\mathbf{x}, \Psi) = \sum_{i=1}^{G} \pi_i \varphi_i(\mathbf{x}; \theta_i), \tag{1}$$

where $\mathbf{x}_{1\times d}$ data vector, the $\pi_i$ are mixing proportions such that $\pi_i > 0$ and $\sum_{i=1}^{G} \pi_i = 1$ and $\varphi_i$ is the component pdf with parameters $\theta_i$. These mixing proportions indicate the probability of any particular data point belonging to cluster $i$. Each of the $G$ components in the mixture model are to uniquely represent a sub population in the data. The case for the uniqueness problem is addressed in section 2.1.1. The finite mixture model is completely specified by the parameters in the parameter set $\Psi = (\pi_1, ...\pi_G, \theta_1, ..., \theta_G)$ and where $\theta_i$ is the parameter set that completely specifies the particular probability density function $\varphi_i(\mathbf{x}; \theta_i)$ known as the mixture components. Various types of restriction can be placed on the component densities, for example $\varphi_i(\mathbf{x}; \theta_i) = \varphi(\mathbf{x}; \theta_i)$ - indicating that the type of distribution is invariant under the change of component label. This leaves only the parameters, of a predefined component probability density function $\varphi(\mathbf{x}; \theta_i)$ (in many cases assumed Gaussian), to be determined. This makes for a much simpler model to estimate.

### 2.1.1 Identifiability

Identifiability is important as it allows unique identification of sub populations in the data which is required to perform meaningful clustering. Each subpopulation, and therefore each cluster, must be able to be uniquely identified with a component of the finite mixture model. A clear overview of indentifiability is given by McLachlan and Peel (2000) section 1.14 and is summarised below.

A parametric family $\zeta(\mathbf{x}, \Psi)$ of density functions is said to be identifiable if it is uniquely determined by distinct values of the parameters $\Psi$. That is to say for a given parameter space $\Omega$ the set

$$\{\zeta(\mathbf{x}, \Psi), \quad \Psi \in \Omega\}$$

where

$$\zeta(\mathbf{x}, \Psi) = \zeta(\mathbf{x}, \Psi^{'}) \; \forall \Psi \in \Omega$$

if and only if

$$\Psi = \Psi^{'}$$

Now, for the identifiability of a mixture of distributions, assume $\zeta(\mathbf{x}, \Psi)$ has two component densities densities $\zeta_i(\mathbf{x}, \theta_i)$ and $\zeta_h(\mathbf{x}, \theta_h)$ that both belong to the same parametric family. Then it will still hold that $\zeta(\mathbf{x}, \Psi) = \zeta(\mathbf{x}, \Psi^{'})$ when component labels $i$ and $h$ are interchanged in $\Psi$. That means, as McLachlan points out, that the class of mixtures is identifiable but $\Psi$ is not. He furthers argues that if all the components belong to the same parametric family (this is so we have that we can meaningfully permute parameter labels), then $\zeta(\mathbf{x}, \Psi)$ is invariant under the $a!$ permutations of the component labels in $\Psi$.

Let

$$\zeta(\mathbf{x}, \Psi) = \sum_{i=1}^{a} \pi_i \zeta_i(\mathbf{x}, \theta_i)$$

and

$$\zeta(\mathbf{x}, \Psi^{'}) = \sum_{i=1}^{a^{'}} \pi_i^{'} \zeta_i(\mathbf{x}, \theta_i^{'})$$

Then the class of finite mixtures is said to be identifiable for $\forall \Psi \in \Omega$ if $\zeta(\mathbf{x}, \Psi) \stackrel{D}{=} \zeta(\mathbf{x}, \Psi^{'})$ if and only if $a = a^{'}$ and we can permute the component labels so that $\pi_i^{'} = \pi_i$ and $\zeta_i(\mathbf{x}, \theta_i^{'}) = \zeta_i(\mathbf{x}, \theta_i)$, where $\stackrel{D}{=}$ implies equality in distribution. The lack of identifiability of $\Psi$ caused by the interchanging of component labels can be easily overcome by ordering the parameters, such as the mixing proportions, as

$$\pi_1 < \pi_2 < ... < \pi_g$$

This constraint should be similarly applied to the other parameters in $\Psi$ in order to overcome the problem of interchanging component labels. When estimating parameters, however, this constraint is relaxed (McLachlan and Peel, 2000).

### 2.1.2 Mixture of Gaussian Distributions

Hereafter the component densities will be assumed to be Gaussian, giving the G-component finite mixture model as

$$f(\mathbf{x}, \Psi) = \sum_{i=1}^{G} \pi_i N(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i), \tag{2}$$

where $\mu_i$ and $\Sigma_i$ are the component mean and component covariance matrix, respectively, of the *ith*

Gaussian component $N(\mathbf{x};\mu_i,\Sigma_i)$ given by

$$N(\mathbf{x};\mu_i,\Sigma_i) = (2\pi)^{-\frac{d}{2}}|\Sigma_i|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\mu_i)\Sigma_i^{-1}(\mathbf{x}-\mu_i)^T} \tag{3}$$

In order to estimate the parameter set $\Psi = (\pi_i, \mu_i, \Sigma_i)$, $i = 1, ..., G$, the EM algorithm (Dempster, et al., 1977) is implemented. The procedure is split in to two steps and is given in the context of model-based clustering; First the expectation step is calculated, where each data point is given a probability of belonging to each component (or cluster) given the current estimate of the model parameters. This quantity is denoted $z_{ij}$ indicating the probability of individual $i$ belonging to component $j$. Here, $z_{ij}$ is taken to be either 1 or 0 depending on whether data point $j$ belongs to component $i$ or not. The data points are therefore assigned to the cluster where it has the highest probability of belonging. After this step the algorithm moves to the maximisation step where the new estimates of the parameters are calculated using the current probability of component membership. This procedure is iterated until some stopping criteria is met. The CEM (conditional expectation maximization) algorithm is outline for the Gaussian finite mixture model:

1. Initialise the parameters associated with the normal distribution. $\mathbf{\Psi} = \mathbf{\Psi}^{old}$

2. Expectation Step: compute the probability of component membership given the current estimation of parameters $P(z_{ij}|\mathbf{x}_j, \mathbf{\Psi}^{old})$

3. Maximisation Step: update the parameter estimations, $\mathbf{\Psi}^{new} = \underset{\mathbf{\Psi}}{argmax}(Q(\mathbf{\Psi}, \mathbf{\Psi}^{old}))$

where

$$Q(\mathbf{\Psi}, \mathbf{\Psi}^{old}) = \sum_{j=1}^{N}\sum_{i=1}^{G} P(z_{ij}|\mathbf{x}_j, \mathbf{\Psi}^{old}) ln(P(\mathbf{x}_j, z_{ij}|\mathbf{\Psi}))$$

4. Compute the log likelihood

The log-likelihood is given as follows

$$\begin{aligned} l(\mathbf{\Psi}) &= \sum_{j=1}^{N}\sum_{i=1}^{G} z_{ij}[ln(\pi_i) - \frac{1}{2}ln|\Sigma_i| - \frac{1}{2}tr(\Sigma_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T)] \\ &= f(\pi_i) + f(\boldsymbol{\mu}_i, \Sigma_i) \end{aligned}$$

Now, from Bayes Theorem

$$\begin{aligned} P(z_{kj}|\mathbf{x}_j, \mathbf{\Psi}^{old}) &= \frac{P(\mathbf{x}_j|z_{kj}, \mathbf{\Psi}^{old})P(z_{kj}|\mathbf{\Psi}^{old})}{P(\mathbf{x}_j|\mathbf{\Psi}^{old})} \\ &= \frac{\pi_k N(\mathbf{x}_j|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_i \pi_i N(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)} \end{aligned}$$

The expected complete log-likelihood is given by:

$$\begin{aligned} Q(\mathbf{\Psi}, \mathbf{\Psi}^{old}) &= \mathrm{E}[\sum_j ln(P(\mathbf{x}_j, z_j|\mathbf{\Psi}))] \\ &= \sum_j \sum_i P(z_{ij}|\mathbf{x}_j, \mathbf{\Psi}^{old}) ln(\pi_i N(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)) \\ &= \sum_j \sum_i P(z_{ij}|\mathbf{x}_j, \mathbf{\Psi}^{old}) ln(\pi_i) + \sum_j \sum_i P(z_{ij}|\mathbf{x}_j, \mathbf{\Psi}^{old}) lnN(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i) \end{aligned}$$

Optimizing the above equation with respect to $\pi_k, \boldsymbol{\mu}_k$ and $\Sigma_k$ giving the following updating equations

$$\pi_k^{new} = \frac{1}{N}\sum_j \frac{\pi_k N(\mathbf{x}_j|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_i \pi_i N(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)}$$

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_j P(z_{kj}|\mathbf{x}_j, \boldsymbol{\Psi}^{old})\mathbf{x}_j}{\sum_j P(z_{kj}|\mathbf{x}_j, \boldsymbol{\Psi}^{old})}$$

$$\Sigma_k^{new} = \frac{\sum_j P(z_{kj}|\mathbf{x}_j, \boldsymbol{\Psi}^{old})(\mathbf{x}_j - \boldsymbol{\mu}_k^{new})(\mathbf{x}_j - \boldsymbol{\mu}_k^{new})^T}{\sum_j P(z_{kj}|\mathbf{x}_j, \boldsymbol{\Psi}^{old})}$$

### 2.1.3 Parsimonious Gaussian Mixture Models

Many parsimonious models have been introduced. In the most general case, a heteroscedastic component covariance matrix is assumed. That is, each covariance matrix, $\Sigma_i$, across the different components are assumed different from each other. M. Varjokallio and M. Kurimo (2007) showed how the heteroscedastic case can be overwhelming when running parameter estimation algorithms. The number of parameters to be estimated are $(G-1) + Gd + Gd(d+1)/2$ of which $Gd(d+1)/2$ are attributed to $\Sigma_i$.

An alternative to the above case is the homoscedastic case whereby it is assumed $\Sigma_i = \Sigma$. In this case $(G-1) + Gd + d(d+1)/2$ parameters are to be estimated, $d(d+1)/2$ of which are attributed to the common covariance matrix $\Sigma$.

Another restriction is proposed by Banfield and Raftery (1993) which is the eigenvalue decomposition of the component covariance matrix

$$\Sigma_i = F_i \Lambda_i F_i^T$$

where $\Lambda_i$ is the diagonal matrix whose entries are the eigenvalues of $\Sigma_i$ and $F_i$ is the matrix of eigenvectors. The orientation of the eigenvectors, or principal components, is determined by the matrix $F_i$, and the eigenvalues in $\Lambda_i$ determine the size and shape of the density contours (Banfield and Raftery, 1993). Furthermore, let $\Lambda_i = \lambda_i A_i$ with matrix $A_i$=$diag\{a_{1i}, ..., a_{pi}\}$ and $1 = a_{1i} \geq ... \geq a_{pi} > 0$. A table of parsimonious models are given below. The Shape refers to the geometrical shape of the cluster. The Orientation describes the restriction placed on each cluster that aligns them in different ways, or together. Size describes the space each cluster occupies in $p$-dimensional space, and whether or not each cluster is the same or different in this regard. Table 4, first shown in Banfield and Raftery (1993) and replicated here for convenience gives parameterisation of the covariance matrix, model name, and other attributes shows the set of parsimonious models.

Table 4: *Eigenvalue decomposition of the component covariance matrix of the Gaussian mixture model*

| $\Sigma_i$ | Name | Size | Shape | Orientation | Number of Parameters |
|---|---|---|---|---|---|
| $\lambda\mathbf{I}$ | EII | Equal | Spherical | None | 1 |
| $\lambda_i\mathbf{I}$ | VII | Different | Spherical | None | $G$ |
| $\Sigma$ | EEI | Equal | Equal | Equal | $d(d+1)/2$ |
| $\lambda_i\Sigma$ | VEI | Different | Equal | Equal | $G + d(d+1)/2$ |
| $\lambda F_i A F_i^T$ | EEV | Equal | Equal | Different | $Gd^2 + d + 1$ |
| $\lambda_i F_i A F_i^T$ | VEV | Different | Equal | Different | $Gd^2 + d + G$ |
| $\lambda_i F A_i F^T$ | EVV | Different | Different | Equal | $d^2 + Gd + G$ |
| $\Sigma_i$ | VVV | Different | Different | Different | $Gd(d+1)/2$ |

Other covariance matrix decompositions are possible. The factor analytic model (Spearman, 1903) places the restriction $\Sigma_i = \mathbf{B}\mathbf{B}^T + \psi$, where the matrix $\mathbf{B}$ is known as the factor loadings, and matrix $\psi$ is known as the individual specific variance. This model is discussed in more detail in section 2.2.

## 2.2 Factor Analysis

This section will introduce the concept of factor analysis and show how a mixture of factor analysers is a suitable way of not only uncovering hidden independent variables (called factors) that can be interpreted to explain the data (Pournara and Wernisch, 2007), but also how it enables the normal mixture model to be fitted to high dimensional data (McLachlan *et al*, 2002). Confirmatory factor analysis confirms or reject a hypothesis about the underlying factors that relate to the data. The researcher uses his knowledge of the theory to postulate the number of factors, and then confirms or rejects this hypothesis with the statistical results of CFA. Exploratory factor analysis seeks to determine the the number of latent factors by analysing the structure of interrelated variables without affecting the structure of the data itself.

One approach to dealing with high dimensional data is to utilise some form of data reduction technique. Principal component analysis is one such method where the observation vectors are projected on to a dimensional plane spanned by the eigenvectors of the data matrix. To reduce the dimensions of the data, we take only the first few eigenvectors that explain most of the variance in the data. This technique, however, does not take in to consideration the task of classification, resulting in poorly classified data (Bouveyron and Brunet, 2012). As such, much of the features of a model are not conserved. These features may hold some value and one may wish maintain the structure of the data while reducing dimension. One such method is factor analysis which can be traced as far back as 1904 (Spearman, 1904). The idea of factor analysis is to reduce the dimensions of the data while keeping the covariance structure of the model the same. Factor analysis can also offer to discover unobservable, or latent, variables in data. This can help the analyst to uncover the reasons, or causes, for the structure of the data. Factor analysis has a wide range of applications in a biological context.

Before the mathematical definition of factor analysis is explained, it is important to have an intuitive idea of what it is. To explain the factor analytic model and how the group structure is laid out, an artificial example as shown in Abdi (2003) will be used. Here he described 5 different wines with 7 variables. The data matrix is given by

| | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Wine 1 | 14 | 7 | 8 | 7 | 7 | 13 | 7 |
| Wine 2 | 10 | 7 | 6 | 4 | 3 | 14 | 7 |
| Wine 3 | 8 | 5 | 5 | 10 | 5 | 12 | 5 |
| Wine 4 | 2 | 4 | 7 | 16 | 7 | 11 | 3 |
| Wine 5 | 6 | 2 | 4 | 13 | 3 | 10 | 3 |

The first two factors are given below with their loadings on each of the seven variables. This is known as the factor loading matrix

| | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Factor 1 | -0.3965 | -0.4454 | -0.2646 | 0.4160 | -0.0485 | -0.4385 | -0.4547 |
| Factor 2 | 0.1149 | -0.1090 | -0.5854 | -0.3111 | -0.7245 | 0.0555 | 0.0865 |

The plane that the data points are project on to are spanned by the vectors labeled "Factor 1" and "Factor 2". Note that there will be as many factors as dimensions of the data, however we choose only the factors that explain most of the data. Furthermore, we are obviously doing no data reduction if we choose the number of factors to be equal to the number of original dimensions. A principal component analysis (PCA) was conducted to unveil a four factor model. These two vectors above are the eigenvectors with corresponding eigenvalues 4.7627 and 1.8101 and were the only two with eigenvalues over 1, and which account for 94%

of the variance. In many cases, the factors have a real world meaning that is applicable to the population. Using the above example, Abdi (2003) finds that the first factor relates to the pricing of the wines, and the second factor relates to their sweetness. This can be seen by how much each variable "loads" on each factor. These loadings represent how much each variable is associated with each factor; "sweetness" (eg "sugar" and "for dessert") and "price". In this way it can be seen that two factors can almost entirely explain the rating of a wine; These two factors are price and sweetness. Furthermore, these factors can be rotated, without the loss of the amount of variance that is explained, by an orthogonal matrix. This is used to interpret the factors more effectively and is covered in section 2.2.3.

To generalise the above example assume $p$-variate data that are projected on to $q$-dimensional hyperplane is given. A factor analytic model can essentially be seen as a calculation of the covariance matrix as if all the variance is explained over a $q$ dimensional plane in $p$-dimensional, for $q < p$. In this sense, the variance and covariance is expected to be explained only in the plane defined by the orthogonal factors, and therefore a restriction on the covariance matrix is imposed to be defined in such a way that the other $p - q$ dimensions have no bearing on it's values. Since these dimensions are effectively dropped from our calculations some estimation error is introduced.

Since there is a projection on to a lower dimensional plane our original data matrix will not be used but rather a matrix of projected vectors (or representative points). This matrix is known as the factor scores matrix. Each projected vector lies on the $q$-dimensional plane that is spanned by the vectors of factors which make up the rows of the factor loading matrix. Without noise (which is associated with the error), the data is expected to lie exactly on this plane. With noise, the data vectors are expected to be scattered around this plane with some error described by Gaussian distribution with mean 0 and covariance matrix $\psi$. The error term is responsible for estimating the information 'lost' when the $p$-dimensional data is projected to a lower dimensional space.

*Remark* : "Principal component analysis and factor analysis are very closely related projection techniques. This similarity can bring about some confusion when being introduced the factor analysis. PCA is purely a mathematical technique that is used to project data on to a subspace which is used as a data reduction technique if the subspace is a lower dimension than the original space. The aim of PCA is to minimise the mean squared distance from the data points to the projections. This, in turn, is done by preserving the variance. It allows for no statistical inferences and does not explain the stochastic processes that influenced the data. Factor analysis, on the other hand, preserves the correlations between variables through the factors. That is to say, the correlation matrix of the projected vectors will be constructed in such a way that it is as close to as possible the correlation matrix of the original feature vectors, while still reducing the number of dimensions. In general the principle components of a data set will not be the same as the factors in the factor model. However, as Tipping *et al* (1999) point out, under certain conditions the principle space and the factor space can be very similar, and in certain cases the principal components are calculated and used as factors in factor analysis.

### 2.2.1 Single Factor Analysis

Let $\mathbf{Y_1}...\mathbf{Y_n}$ denote a set of $n$ $p$-dimensional observations. The factor analysis model for the data matrix $\mathbf{Y}$ is

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{BU} + \epsilon \tag{4}$$

$\mathbf{Y}$ is the $p \times n$ data matrix $\boldsymbol{\mu}$ $p \times 1$ the mean. $\mathbf{B}$ is the $p \times q$ matrix of factor loading ($q \leq p$) , $\mathbf{U}$ is the $q \times n$ matrix of factor scores and $\epsilon$ is the error matrix. The factor scores $\mathbf{U}$ are assumed independent and identically distributed (i.i.d) as $N(0, \mathbf{I}_q)$, independent of $\epsilon$ which is i.i.d as $N(0,\psi)$ where $\psi$ is generally assumed to be a diagonal matrix. Each individual observation in the data matrix is i.i.d$\sim N(\mu + \mathbf{BU},\psi)$, if

the matrix $\mathbf{U}$ would be known (conditional on $\mathbf{U}$). Unconditionally, $\mathbf{Y} \sim N(\mu, \mathbf{B\Upsilon B}^{\mathrm{T}} + \psi)$ (see A.2), and where the simplifying assumption $\Upsilon = \mathbf{I}$ is made.

### 2.2.2 Identifiability, Constraints and Effective Number of Free Parameters

Constraints are usually needed in order to make the factor model identifiable. To see this suppose a parameterisation of the covariance matrix is given as

$$\underset{(\mathrm{p}\times\mathrm{p})}{\mathbf{\Sigma}} = \underset{(\mathrm{p}\times\mathrm{q})}{\mathbf{B}}\underset{(\mathrm{q}\times\mathrm{p})}{\mathbf{B^T}} + \underset{(\mathrm{p}\times\mathrm{p})}{\psi} \ . \tag{5}$$

The left hand side has $p(p+1)/2$ free parameters (since the covariance matrix $\Sigma$ is symmetric), where as the right hand side has $pq + p(p+1)/2$. The problem here is that, in general, if there are more equations (degrees of freedom of $\Sigma$) than unknowns (total degrees of freedom of $\mathbf{B}\mathbf{B}^T + \psi$) then there is generally no solution. If there are more unknowns than equations then there are infinitely many solutions, making the factor model unidentifiable since infinitely many different factor models can represent the same covariance matrix. There are, however, several ways to reduce the number of these parameters. Firstly, the observable variables can be scaled to have a standard deviation of 1. This means that there are now $p(p-1)/2$ degrees of freedom in $\Sigma$. That is to say there are $p(p-1)/2$ equations on the left hand side. On the right hand side of equation 5 the restriction that $\psi$ be diagonal is made. This assumption is made to ensure that whatever correlations occur, they occur only though the factors. This means that there are $p$ unknowns in $\psi$ to solve. Since $\Sigma$ has been standarised as mentioned, the diagonal elements of $\psi$ are known as soon as the elements of $\mathbf{B}$ are known. Furthermore, following the reasoning from Lawley and Maxwell (1971) a criterion that the matrix $\mathbf{B}^T\psi^{-1}\mathbf{B}$ be diagonal must be met, which results in $\frac{1}{2}q(q-1)$ constraints on parameters of $\mathbf{B}$. This leaves $p(p-1)/2$ equations and $pq - \frac{1}{2}q(q-1)$ unknowns. In order to get a unique solution set $q = p$, but then no data reduction is being done. If $q$ is too large relative to $p$ then the model is unidentifiable. If $q$ is too small the system is overdetermined which can only be satisfied for certain restricted $\Sigma$. Henceforth it will be assume that $\psi$ is strictly diagonal.

### 2.2.3 Orthogonal Factor Rotations

The factors extracted by performing factor analysis on a data set are generally constructed in a way that they are orthogonal to each other. In practice, only a portion of the numbers of factors are kept. The remaining factors are discarded, or assumed to be noise or measurement error (H. Abdi, 2003). These factors are rotated in order to infer a suitable interpretation of their meaning relative to the data at hand. This can be done through an orthogonal transformation. To see this, suppose the data matrix to be given by $\mathbf{Y}$ and some orthogonal matrix $\mathbf{Q}$. Then, since $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$

$$\begin{aligned} \mathbf{Y} &= \mathbf{UB} + \epsilon \\ &= \mathbf{UQQ^TB} + \epsilon \\ &= \mathbf{U^*B^*} + \epsilon \end{aligned}$$

where $\mathbf{U}^* = \mathbf{UQ}$ and $\mathbf{B}^* = \mathbf{Q}^T\mathbf{B}$. This demonstrates that a rotation of factors in the sub-space does not affect the feature of the original matrix. Factor rotations are used to maximise the loadings of certain variables on to their respective factors, while reduce the loadings of other variables on these same factors. This way there are fewer variables with higher loadings on the few factors, than many variables all with average loading values. In this way a factor can be more accurately interpreted in terms of a real world context. However, these new factors do not necessarily produce better results. They merely produce more interpretable results. To use the example of wine data, H. Abdi (2003) showed the factor loading matrix

after it was rotated clockwise by 15 degrees, giving

| | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Factor 1 | -0.4124 | -0.4057 | -0.1147 | 0.4790 | -0.1286 | -0.4389 | -0.4620 |
| Factor 2 | 0.0153 | -0.2138 | -0.6321 | -0.2010 | -0.7146 | 0.0525 | 0.0264 |

The rotations did not show much change for the interpretation of the first dimension (Factor 1), but Factor 2 now appears more clearly to be a dimension of sweetness.

### 2.2.4 Mixture of Factor Analysers

Single factor analysis is extremely useful, but in certain cases can be quite limiting by its global linearity (McLachlan and Peel, 2000). This means that while a portion of the data can be explained by a hyperplane spanned by a set of vectors (known as the factors) in d-dimensional space (d-variate data), it is conceivable, and indeed likely, that not all the data in the d-space can be explained by this single hyperplane. The alternative is to create another plane of equal dimension to the previous, spanned by different vectors, that exists to explain the data in a different region of space. As McLachlan and Peel point out, a non-linear model of factor analysers can be formulated by summing over a finite mixture of linear sub model for the full observation matrix $\mathbf{Y}$, given the matrix of factors $\mathbf{U}$. That is to say that one can reduce dimensionality, and in turn reduce parameters, by modeling $\mathbf{Y}$ as

$$\mathbf{Y}_{|C_i} = \boldsymbol{\mu}_i + \mathbf{U}_i\mathbf{B}_i + \epsilon_i \tag{6}$$

with associated prior probabilities $\pi_i$ $(i = 1, \ldots, g)$, $\epsilon_{\mathrm{i}} \sim N(0, \psi_{\mathrm{i}})$ and cluster $C_i$.

To reduce the number of parameters of the finite mixture model is to reduce the dimensions in the variable space to a subspace of factors and find the associated parameters in this reduced factor subspace. That is to say, for any finite mixture model

$$f(\mathbf{x}; \boldsymbol{\Psi}) = \sum_{i=1}^{G} \pi_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{7}$$

where

$$\boldsymbol{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \psi_{\mathrm{i}} \tag{8}$$

The parameter vector $\boldsymbol{\Psi}$ now contains the parameters of $\boldsymbol{\mu}_i$, $\mathbf{B}_i$ and $\psi_i$ and are to be estimated. There are 12 parsimonious models proposed by P. McNicholas and T. Murphy (2010) and are given in the Table 5. Note that $\boldsymbol{\Delta}_i$ is a diagonal matrix such that $|\boldsymbol{\Delta}_i| = 1$ meaning that the product of the diagonal elements must be 1. Also, assume $\sigma_i \in +\mathbb{R}$ and where $\mathbf{I}_d$ is the $d \times d$ identity matrix.

Table 5: *Parsimonious Gaussian subspace models*

| Model | $\mathbf{B}_i = \mathbf{B}$ | $\mathbf{\Delta}_i = \mathbf{\Delta}$ | $\sigma_i = \sigma$ | $\Delta_i = \mathbf{I}_d$ | Covariance Matrix |
|---|---|---|---|---|---|
| CCCC | Constrained | Constrained | Constrained | Constrained | $\mathbf{\Sigma}_i = \mathbf{B}\mathbf{B}^T + \sigma\mathbf{I}_p$ |
| CCUU | Constrained | Constrained | Unconstrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}\mathbf{B}^T + \sigma_i\mathbf{\Delta}$ |
| CUCU | Constrained | Unconstrained | Constrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}\mathbf{B}^T + \sigma\mathbf{\Delta}_i$ |
| CUUU | Constrained | Unconstrained | Unconstrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}\mathbf{B}^T + \sigma_i\mathbf{\Delta}_i$ |
| UCCC | Unconstrained | Constrained | Constrained | Constrained | $\mathbf{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma\mathbf{I}_p$ |
| UCUC | Unconstrained | Constrained | Unconstrained | Constrained | $\mathbf{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma_i\mathbf{I}_p$ |
| UUCU | Unconstrained | Unconstrained | Constrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma\Delta_i$ |
| UUUU | Unconstrained | Unconstrained | Unconstrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma_i\Delta_i$ |
| CCCU | Constrained | Constrained | Constrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}\mathbf{B}^T + \sigma\Delta$ |
| UCCU | Unconstrained | Constrained | Constrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma\Delta$ |
| UCUU | Unconstrained | Constrained | Unconstrained | Unconstrained | $\mathbf{\Sigma}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma_i\Delta$ |
| CCUC | Constrained | Constrained | Unconstrained | Constrained | $\mathbf{\Sigma}_i = \mathbf{B}\mathbf{B}^T + \sigma_i\mathbf{I}_p$ |

These parsimonious models further reduce the number of parameters of the model of a mixture of factor analysers. Ghaharamani and Hinton (1997) assume the UCCU model in their analysis whereas McLachlan and Peel (2000) assume the UUUU model. The UCCC, CCUC, UCUC and CCCC cases could all be assumed for probabilistic principle component analysis shown by Tipping and Bishop (1999).

## 2.3 Simultaneous Clustering with a Standard Mixture Model

The aim of simultaneous clustering is to find a statistical link between the parameters of the models of two different populations and use this to find the parameters of the statistical model of the unclassified sample. Firstly, it has been shown (Appendix A.1) that a linear map can be found that gives a stochastic link between two different non-degenerate, univariate normal distributions transforming one, in distribution, in to the other. Lourme and Biernacki (2012) showed that this can be utilised when transforming, in distribution, one mixture model in to another, giving birth to the so-called simultaneous clustering method, and allowing the user to find the linear stochastic link between two populations. Finding a link between two populations has been done in various contexts; Biernacki et el (2002); Van Franeker and Ter Brack (1993); Lourme and Biernacki, (2010). In certain situations several samples arising from different populations need to be clustered. It is assumed a similarity exists between these populations so that simultaneous model-based clustering can be performed. It is also assumed that each population have the same number of subgroups, where each subgroup has the same meaning, and that each population be described by the same meaning variables. One can either cluster each sample independently, or use the method of simultaneous model based clustering (Lourme and Biernacki, 2012). At face value this allows the user of the proposed method to subsequently cluster a second sample without estimating the usual parameters of the mixture model, but rather estimate the parameter of the stochastic function that links the previously clustered sample and the new one. This provides several benefits. Firstly, it allows the practitioner to simultaneously cluster several samples between these different populations. Secondly, it stops the redundancy of having to cluster several samples when it is reasonable to assume a common underlying group structure. More precisely, the M-step in the GEM algorithm presented in Lourme and Biernacki (2012) need only maximise with respect to the parameters found in the first population and the parameters in the stochastic functions linking the first with the $h$th population. The process is as follows; Firstly, it is assumed that one sample is measured, and the parameters associated with modeling the population are estimated in the usual way. This is the reference sample. A link between the two populations, the reference and non-reference population, is to be found and in turn estimate the parameters of the other unknown population without actually estimating its parameters directly. This leads to a considerable reduction in the number of parameters estimated. The core differences between simultaneous clustering with a mixture of normal distributions and independent

clustering are summarised in the following 3 points;

- Firstly, simultaneous clustering allows one clustering procedure to cluster all of the multiple sample of the different, but similar, populations by assuming a statistical link between each population. Independent clustering requires a separate procedure for each sample of each population and assumes no statistical links.
- Secondly, simultaneous clustering has some non-allowed parsimonious models. The non-allowed combination of interpopulation and intrapopulation models are caused by nonsensical restrictions placed on these parameters of the model. Independent clustering has no such non-allowed models due to there being no interpopulation parameters.
- Thirdly, the samples, in the case for simultaneous clustering, are not directly clustered with exception of the reference sample. Instead, the parameters of the link functions are estimated and it is with these that the model parameters of the remaining samples are estimated. With independent clustering, the parameters of model for each sample are estimated individually.

### 2.3.1 The Data

The data consists of several samples. Suppose sample $S^h$ (measured and known) is taken from population $P^h$ and $S^{h'}$, different from $S^h$, taken from population $P^{h'}$ where $P^h$ and $P^{h'}$ need not necessarily be the same population.

$S^h$ will be modeled by a finite mixture model with all parameters estimated. Each of the $n^h$ observations of sample $S^h$ is described by an ordered pair $(\mathbf{x}_i^h, z_{ij}^h)$, $1 < i < n^h$ and $1 < j < K$, where $\mathbf{x}_i^h$ is the observed vector of $d$ variables on the $i$th realisation, and $z_{ij}$ is the indicator variable that is 1 if the $i$th observation belongs to the $j$th cluster of the mixture model, and 0 otherwise. Each of these pairs of observations, $(\mathbf{x}_i^h, z_{ij}^h)$, are realisations of the random variable $(X^h, Z^h)$ where

$$(X^h | Z_j^h = 1) \sim N(\mathbf{x}^h; \boldsymbol{\mu}_j^h, \Sigma_j^h)$$

and

$$Z^h \sim B_K(p_1^h, ..., p_K^h)$$

where $B_k$ is the Bernoulli distribution with parameters $p_1^h, ..., p_K^h$ with $p_k^h$ the proportion of the $k$th group in the population $P^h$. $N(\mathbf{x}^h; \boldsymbol{\mu}_j^h, \Sigma_j^h)$ is the $d$ dimensional normal distribution with the mean $\boldsymbol{\mu}_j^h \in \mathbb{R}^d$ and covariance matrix $\Sigma_j^h \in \mathbb{R}^{d \times d}$. It is assumed that the above parameters ,$\Psi^h = \{z_{ij}^h, \boldsymbol{\mu}_j^h, \Sigma_j^h\}$, are known or calculated with the appropriate numerical methods. This is known as our reference sample.

Given sample $S^{h'}$ with $n^{h'}$ observations on it where only the $\mathbf{x}_i^{h'}$'s are known. Each observation of the sample $S^{h'}$ is described by the ordered pair $(\mathbf{x}_i^{h'}, z_{ij}^{h'})$,$1 < i < n^{h'}$ and $1 < j < K$, with the entire parameter space, $\Psi^{h'} = \{z_{ij}^{h'}, \boldsymbol{\mu}_j^{h'}, \Sigma_j^{h'}\}$, unknown. The pair $(\mathbf{x}_i^{h'}, z_{ij}^{h'})$ are realisations of the random variable $(X^{h'}, Z^{h'})$ modeled as follows

$$(X^{h'} | Z_j^{h'} = 1) \sim N(\mathbf{x}^{h'}; \boldsymbol{\mu}_j^{h'}, \Sigma_j^{h'})$$

and

$$Z^{h'} \sim B_K(p_1^{h'}, ..., p_K^{h'})$$

The aim is to find the parameters in $\Psi^{h'}$ using $S^h$ and $S^{h'}$, by finding a link between the populations $P^h$ and $P^{h'}$.

### 2.3.2 Simultaneous model-based clustering and the linear stochastic link

Simultaneous clustering is applied to a finite mixture model of Gaussians as follows; Any distribution can be approximated arbitrarily close by a mixture of G distributions. Firstly, the aim is to separate H samples in to these G groups. Each sample $h$,$h \in (1,\ldots,H)$ , is composed of $n^h$ individuals $\mathbf{x}_{\mathrm{i}}^{\mathrm{h}}$, $\mathrm{i}=1,\ldots,\mathrm{n}^{\mathrm{h}}$, of $\mathbb{R}^d$. The density function is given as follows:

$$f(\mathbf{x};\mathbf{\Psi}^{\mathrm{h}}) = \sum_{i=1}^{G} \pi_i^h N(\mathbf{x};\boldsymbol{\mu}_i^h,\Sigma_i^h), \qquad \mathbf{x} \in \mathbb{R}^d$$

The coefficient $\pi_i^h$, $i = 1,\ldots,G$, are the mixing proportions of sample $h$, $\boldsymbol{\mu}_i^h$ refers to the center of component $i$ of population sample $h$, and $\Sigma_i^h$ refers to the component co-variance matrix of population $h$.

The solution is shown by A. Lourme and C. Biernacki (2010). Firstly, the subgroups to be discovered consist of the same features, and same meaning partitions in each sample. It is with this assumption that a distributional relationship is assumed to exist between two different samples.

The form of the stochastic link is to be known before the parameters of the link function can be practically estimated, and applied. It has been found that any mapping between two Gaussians is linear ( Biernacki et al 2002) and, under a certain transformation, are equal in distribution. Hence there exists a matrix $\mathbf{D}_i^{h,h'} \in \mathbb{R}^{d\times d}$ diagonal and $\mathbf{b}_i^{h,h'} \in \mathbb{R}^d$ such that:

$$(\mathbf{X}^{h'}|Z_i^{h'}=1) = \mathbf{D}_i^{h,h'}(\mathbf{X}^h|Z_i^h=1) + \mathbf{b}_i^{h,h'} \tag{9}$$

where $z_j^h \in \{0,1\}^G$ $(j=1,...,n^h)$ is the unobserved group data and indicates component membership such that $z_{ij}^h = 1$ if $\mathbf{x}_i^h$ arose from component $C_i^h$ and 0 otherwise.

Thus, the covariance and mean, respectively, for each $i$ and $h'$ are calculated as follows: The result will be proved with the use of the moment generating function of the multivariate normal distribution. For any given component membership $k$ the moment generating function of the multivariate normal distribution is

$$M_{\mathbf{X}_i^h}(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}_i^h + \frac{1}{2}\mathbf{t}\Sigma_i^h\mathbf{t}}$$

where $\mathbf{X}_i^h$ is shorthand for $(\mathbf{X}^h|Z_i^h=1)$. Similarly with $\mathbf{X}_i^{h'}$. Then

$$\begin{aligned}
M_{\mathbf{X}_i^{h'}}(\mathbf{t}) &= E[e^{\mathbf{t}^T\mathbf{X}_i^{h'}}] \\
&= E[e^{\mathbf{t}^T(\mathbf{D}_i^{h,h'}\mathbf{X}_i^h + \mathbf{b}_i^{h,h'})}] \\
&= E[e^{\mathbf{t}^T\mathbf{D}_i^{h,h'}\mathbf{X}_i^h} e^{\mathbf{t}^T\mathbf{b}_i^{h,h'}}] \\
&= e^{\mathbf{t}^T\mathbf{b}_i^{h,h'}} E[e^{((\mathbf{D}_i^{h,h'})^T\mathbf{t})^T\mathbf{X}_i^h}] \\
&= e^{\mathbf{t}^T\mathbf{b}_i^{h,h'}} M_{\mathbf{X}_i^h}((\mathbf{D}_i^{h,h'})^T\mathbf{t}) \\
&= e^{\mathbf{t}^T\mathbf{b}_i^{h,h'}} e^{((\mathbf{D}_i^{h,h'})^T\mathbf{t})^T + \frac{1}{2}((\mathbf{D}_i^{h,h'})^T\mathbf{t})^T\Sigma_i^h(\mathbf{D}_i^{h,h'})^T\mathbf{t}} \\
&= e^{\mathbf{t}^T(\mathbf{D}_i^{h,h'}\boldsymbol{\mu}_i^h + \mathbf{b}_i^{h,h'}) + \frac{1}{2}\mathbf{t}^T\mathbf{D}_i^{h,h'}\Sigma_i^h(\mathbf{D}_i^{h,h'})^T\mathbf{t}}
\end{aligned}$$

The moment generating function is unique to a specific distribution. Also, since $\mathbf{D}_i^{h,h'}\Sigma_i^h\mathbf{D}_i^{h,h'}$ is positive-definite and $> 0$ then $(\mathbf{X}^{h'}|Z_i^{h'}=1) \sim N(\boldsymbol{\mu}_i^{h'},\Sigma_i^{h'})$ where

$$\Sigma_i^{h'} = \mathbf{D}_i^{h,h'}\Sigma_i^h\mathbf{D}_i^{h,h'} \tag{10}$$

and

$$\boldsymbol{\mu}_i^{h'} = \mathbf{D}_i^{h,h'} \boldsymbol{\mu}_i^h + \mathbf{b}_i^{h,h'} \tag{11}$$

This mapping allows the user to estimate the mean and covariance structure of all populations by estimating the link parameters $\mathbf{D}_i^{h,h'}$ and $\mathbf{b}_i^{h,h'}$ from the reference population.

To get an intuitive sense of what the parameters $\mathbf{D}_i^{h,h'}$ and $\mathbf{b}_i^{h,h'}$ are, the following should be noted. The parameter $\mathbf{b}_i^{h,h'}$ represent the 'shift' of each cluster in the variable space. To see this refer to equation 11. It is clear that the mean vector of the population $h$, which could be rephrased as the centre of cluster $h$, is being shifted by the vector $\mathbf{b}_i^{h,h'}$. The matrix $\mathbf{D}_i^{h,h'}$ represents the scaling of the size and shape of the cluster in the variable space. Equation 10 shows how the change in the covariance matrix of cluster $h$, which describes how 'spread out' the cluster is in the variable space, is being affected by the matrix $\mathbf{D}_i^{h,h'}$.

### 2.3.3 Parsimonious Models

Following Lourme and Biernacki (2012) some parsimonious models are introduced. This involves combining assumptions made about each component of the Gaussian mixture models (intrapopulation models) with assumptions about equation 9 (interpopulation models).

Some Intrapopulation models have been discussed above where constraints of the covariance matrices were given. Some other constraints can be enforced, such as $\pi_i^h = \pi^h$. Supplementary to these constraints are the interpopulation constraints whereby the parameters $\mathbf{D}_i^{h,h'}$ and $\mathbf{b}_i^{h,h'}$ are restricted. The most general case is where the matrix $\mathbf{D}_i^{h,h'}$ is positive definite and diagonal, and $\mathbf{b}_i^{h,h'}$ unconstrained. Furthermore, the constraint $\pi_i^h = \pi_i$ could be considered. This constraint depends on whether the ratios $\alpha_i^{h,h'} = \frac{\pi_i^{h'}}{\pi_i^h}$ equals 1 or not (Lourme and Biernacki, 2012). Certain combination of inter- and intrapopulation restriction are not identifiable or allowed. See Lourme and Biernacki (2012) for the models pertaining to the restriction $\Sigma_i^h = \Sigma^h$ and for the general case $\Sigma_i^h$ for a list of these combinations. The parsimonious models proposed in Table 1 were not considered in Lourme and Biernacki (2012), nor in any other literature since then. To identify the non-allowed models pertaining to the intrapopulation model restriction of Table 1 one would have to pay attention to equations 10 and 11. It would be impossible, for example, to assume the model has a different eigenvector decomposition between groups but assume that the link function between samples is group dependent. Therefore the model $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi^h, \lambda_i^h, \mathbf{F}_i^h, A_i^h\}$ would not be allowed. Furthermore, since the matrix $\mathbf{D}_i^{h,h'}$ offers to rotate the random vector in the variable space it would make no sense to allow the model to force each cluster to have equal orientation and then allow this transformation to be component-dependent. The model $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi^h, \lambda_i^h, \mathbf{F}^h, A_i^h\}$ would therefore not be allowed. Similarly, since $\mathbf{D}_i^{h,h'}$ also offers to rescale the random vector, keeping free orientations and sizes and equal shapes, the model $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi^h, \lambda_i^h, \mathbf{F}_i^h, A^h\}$ would not be allowed. Table 3 gives all the allowed and non-allowed models.

Table 6: *Parsimonious models for simultaneous mixture of Gaussian distributions. "•" represents allowable models and "·" non-allowable models*

| | | | $\pi_i$ - $\pi$ | | | | |
|---|---|---|---|---|---|---|---|
| | | | $\lambda^h\mathbf{I}$ | $\lambda_i^h\mathbf{I}$ | $\lambda_i^h F_i^h A^h (F_i^h)^h$ | $\lambda_i^h F^h A_i^h (F^h)^T$ | $\lambda^h F_i^h A^h (F_i^h)^T$ |
| $\pi(\pi^h)$ | $\mathbf{I}$,$\alpha^{h,h'}\mathbf{I}$,$\mathbf{D}^{h,h'}$ | $\mathbf{0}$ | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) |
| | | $\mathbf{b}^{h,h'}$ | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) |
| | | $\mathbf{b}_i^{h,h'}$ | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) | • (•) - • (·) |
| | $\alpha_i^{h,h'}\mathbf{I}$,$\mathbf{D}_i^{h,h'}$ | $\mathbf{0}$ | · (·) - · (·) | • (•) - • (·) | ·(·) - · (·) | ·(·) - ·(·) | ·(·) - · (·) |
| | | $\mathbf{b}_i^{h,h'}$ | · (·) - · (·) | • (•) - • (·) | ·(·) - · (·) | ·(· ) - ·(·) | ·(·) - · (·) |

## 2.4 Overlapping of Clusters

The importance of the overlapping of clusters is notable when dealing with the error rate associated with clustering algorithms. The overlap of a cluster, in the intuitive sense, is simply when the data points of different components associated with its particular cluster are found within the region that is defined by another cluster with associated component. The strict definition of overlap given in Maitra and Malnykov (2010) is used in the 'MixSim' method used in section 4. The results therein demonstrate how an increase in overlap can cause the misclassification rate to increase.

Overlap is defined as follows; Suppose $g(x) = \sum_{i=1}^{g} \pi_i \phi(\mathbf{X}, \boldsymbol{\mu}_i, \Sigma_i)$ is the finite mixture model that describes a population, and $\phi(\mathbf{X}, \boldsymbol{\mu}_k, \Sigma_k)$ is the specific component that models cluster k. The definition of overlap $\omega_{ij}$ (Maitra and Malnykov, 2010) is given by the sum of the two misclassification probabilities

$$\omega_{j|i} = P[\pi_i \phi(\mathbf{X}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) < \pi_j \phi(\mathbf{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) | \mathbf{X} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)] \tag{12}$$

$$\omega_{i|j} = P[\pi_j \phi(\mathbf{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) < \pi_i \phi(\mathbf{X}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) | \mathbf{X} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)] \tag{13}$$

$\phi(\mathbf{X}, \boldsymbol{\mu}_i, \Sigma_i)$ refers to the cluster that is defined by centre $\boldsymbol{\mu}_i$ and covariance $\Sigma_i$. The misclassification probability $\omega_{j|i}$ refers to the probability that the $d$-dimensional data point $\mathbf{x}$ that belongs to cluster $\phi(\mathbf{X}, \boldsymbol{\mu}_i, \Sigma_i)$ be misclassified to belong to the cluster that is defined by the component $\phi(\mathbf{X}, \boldsymbol{\mu}_j, \Sigma_j)$. The overlap is therefore defined as $\omega_{ij} = \omega_{i|j} + \omega_{j|i}$.

## 2.5 Parameter Estimation Via the EM Algorithm

The parameter estimation technique is done via the EM algorithm (Dempster et al, 1977) or some variation thereof. The overview of the EM algorithm was presented in section 2.1.2 for the Gaussian mixture. In this section, the EM algorithm is presented for the single factor model, mixture of factor analysers, and the SMBC model.

### 2.5.1 Likelihood and Estimating Equations of the Factor Analytic Model

(Ghahramani and Hinton, 1997). Given the observed data set $\mathcal{D} = \{\mathbf{x}_j\}$, factor loading matrix $\mathbf{B}$, latent factors $\mathbf{u}$, mean of observed variables $\boldsymbol{\mu}$, and Gaussian noise $\epsilon$ with distribution $\mathcal{N}(0, \psi)$. The aim is to estimate the parameters in $\mathbf{B}$ and $\psi$. The complete data log-likelihood is given by

$$
\begin{aligned}
L_c(\mathbf{B},\psi) &= ln\prod_{j=1}^{N} P(\mathbf{x}_j,\mathbf{u}_j|\mathbf{B},\psi) \\
&= ln\prod_{j=1}^{N} P(\mathbf{x}_j,\mathbf{u}_j|\mathbf{B},\psi) \\
&= \sum_{j=1}^{N} ln(P(\mathbf{x}_j,\mathbf{u}_j|\mathbf{B},\psi)) \\
&= \sum_{j=1}^{N} ln(P(\mathbf{x}_j|\mathbf{u}_j,\mathbf{B},\psi)P(\mathbf{x}_j|\mathbf{B},\psi) \\
&= \sum_{j=1}^{N} ln(P(\mathbf{x}_j|\mathbf{u}_j,\mathbf{B},\psi) + \sum_{j=1}^{N} ln(P(\mathbf{u}_j))
\end{aligned}
$$

where, in the second term of the last line, the restriction that the factor scores are independent of the factor loadings and the individual-specific variance is asserted. Also note that the second summation term goes to zero when maximising the likelihood, so the likelihood function becomes

$$
L_c(\mathbf{B},\psi) = \sum_{j=1}^{N} ln(P(\mathbf{x}_j|\mathbf{u}_j,\mathbf{B},\psi)
$$

The conditional mean and covariance have been calculated above, and hence expanding the likelihood gives

$$
\begin{aligned}
L_c(\mathbf{B},\psi) &= \sum_{j}^{N} ln(\frac{1}{(2\pi)^{\frac{d}{2}}|\psi|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{x}_j - \mathbf{B}\mathbf{u}_j)^T\psi^{-1}(\mathbf{x}_j - \mathbf{B}\mathbf{u}_j))) \\
&= -\frac{Nd}{2}ln(2\pi) - \frac{N}{2}ln|\psi| - \frac{1}{2}\sum_{j}^{N}(\mathbf{x}_j^T\psi^{-1}\mathbf{x}_j - 2\mathbf{x}_j^T\psi^{-1}\mathbf{B}\mathbf{u}_j + tr(\mathbf{B}^T\psi^{-1}\mathbf{B}\mathbf{u}_j\mathbf{u}_j^T))
\end{aligned}
$$

Taking the expectation

$$
\mathrm{E}[L_c] = k - \frac{N}{2}ln|\boldsymbol{\psi}| - \frac{1}{2}\sum_{j}^{N}(\mathbf{x}_j^T\psi^{-1}\mathbf{x}_j - 2\mathbf{x}_j^T\psi^{-1}\mathbf{B}\mathrm{E}[\mathbf{u}_j|\mathbf{x}_j] + tr(\mathbf{B}^T\psi^{-1}\mathbf{B}\mathrm{E}[\mathbf{u}_j\mathbf{u}_j^T|\mathbf{x}_j]))
$$

Maximising with respect to $\mathbf{B}$ and $\psi$ updating equations are given by

$$
\begin{aligned}
\mathbf{B} &= (\sum_{j}^{N}\mathbf{x}_j\mathrm{E}[\mathbf{u}_j|\mathbf{x}_j])(\sum_{j}^{N}\mathbf{x}_j\mathrm{E}[\mathbf{u}_j\mathbf{u}_j^T|\mathbf{x}_j]))^{-1} \\
\psi &= \frac{1}{N}diag[\sum_{j}^{N}\mathbf{x}_j\mathbf{x}_j^T - (\sum_{j}^{N}\mathbf{x}_j\mathrm{E}[\mathbf{u}_j|\mathbf{x}_j]^T)\mathbf{B}^T]
\end{aligned}
$$

The sufficient statistics $\mathrm{E}[\mathbf{u}_j|\mathbf{x}_j]$ and $\mathrm{E}[\mathbf{u}_j\mathbf{u}_j^T|\mathbf{x}_j]$ are given below with their derivations omitted

$$
\begin{aligned}
&\mathrm{E}[\mathbf{u}|\mathbf{x}] = (\mathbf{I} + \mathbf{B}^T\psi^{-1}\mathbf{B})^{-1}\mathbf{B}^T\psi^{-1}\mathbf{x} = \beta\mathbf{x} \\
&\mathrm{E}[\mathbf{u}\mathbf{u}^T|\mathbf{x}] = \mathbf{I} - \beta\mathbf{B} + \beta\mathbf{x}\mathbf{x}^T\beta^T
\end{aligned}
$$

where, for ease of notation define $\beta = (\mathbf{I} + \mathbf{B}^T\psi^{-1}\mathbf{B})^{-1}\mathbf{B}^T\psi^{-1}$

### 2.5.2 Parameter Estimation of the Mixture of Factor Analysers Model

This section will describe the Alternating Expectation Conditional Maximization algorithm as shown in McLachlan et al. (2002). The AECM algorithm is a variant of the EM algorithm. We apply the AECM algorithm to estimate parameters of the MFA model. Firstly the vector of unknown parameters $\mathbf{\Psi}$ is split in to two subvectors $\mathbf{\Psi} = \{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}$ where $\mathbf{\Psi}_1$ contains the mixing proportions $\pi_i \, (i = 1, ..., G)$ and component means $\boldsymbol{\mu}_i \, (i = 1, ..., G)$, and $\mathbf{\Psi}_2$ contains the factor loading matrix $\mathbf{B}_i \, (i = 1, ..., G)$ and individual-specific variances $\psi_i \, (i = 1, ..., G)$. On the first cycle, $\mathbf{\Psi}_1$ is updated and the missing data is considered to be $z_{ij}$ for $i = 1, ..., G$ and $j = 1, ..., n$.

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k)})$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k)})\mathbf{x}_j}{\sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k)})}$$

where,

$$\tau_i(\mathbf{x}_j|\mathbf{\Psi}) = \frac{\pi_i \phi(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)}{\sum_{h=1}^{G} \pi_h \phi(\mathbf{x}_j|\boldsymbol{\mu}_h, \Sigma_h)}$$

On the second cycle $\mathbf{\Psi}_2$ is updated, and the missing data is considered to be $z_{ij}$ for $i = 1, ..., G$ and $j = 1, ..., n$ and the factors $\mathbf{u}_1, ..., \mathbf{u}_n$. When computing the updates, the parameters of $\mathbf{\Psi}_1^{(k+1)}$ are used. Define $\mathbf{\Psi}_1^{(k+\frac{1}{2})} = \{\mathbf{\Psi}_1^{(k+1)}, \mathbf{\Psi}_2^{(k)}\}$. The update formulae are as follows

$$\mathbf{B}_i^{(k+1)} = \mathbf{V}_i^{(k+\frac{1}{2})} \Gamma_i^{(k)} (\Gamma_i^{(k)} \mathbf{V}_i^{(k+\frac{1}{2})} \Gamma_i^{(k)} + \boldsymbol{\omega}_i^{(k)})^{-1}$$

where,

$$\mathbf{V}_i^{(k+\frac{1}{2})} = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k+\frac{1}{2})})(\mathbf{x}_j - \boldsymbol{\mu}_i^{(k+1)})(\mathbf{x}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k+\frac{1}{2})})}$$

$$\Gamma_i^{(k)} = (\mathbf{B}_i^{(k)} \mathbf{B}_i^{(k)^T} + \psi_i)^{-1} \mathbf{B}_i^{(k)}$$

and,

$$\boldsymbol{\omega}_i^{(k)} = \mathbf{I}_q - \Gamma_i^{(k)^T} \mathbf{B}_i^{(k)}$$

The estimates for $\psi_i$ are given by

$$\psi_i^{(k+1)} = diag\{\mathbf{V}_i^{(k+\frac{1}{2})} - \mathbf{B}_i^{(k+1)} \mathbf{H}^{(k+\frac{1}{2})} \mathbf{B}_i^{(k+1)^T}\}$$

where,

$$\mathbf{H}^{(k+\frac{1}{2})} = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k+\frac{1}{2})}) E_i^{(k+\frac{1}{2})}(\mathbf{U}_j \mathbf{U}_j^T|\mathbf{x}_j)}{\sum_{j=1}^{n} \tau_i(\mathbf{x}_j|\mathbf{\Psi}^{(k+\frac{1}{2})})}$$

where $E_i^{(k+\frac{1}{2})}$ denotes the conditional expectation given membership of the $i$th component using $\mathbf{\Psi}^{(k+\frac{1}{2})}$ for $\Psi$.

### 2.5.3 Parameter Estimation of the Simultaneous Model.

Following Lourme and Biernacki (2012) and their use of the GEM algorithm devised by Dempster*et el* (1977), the complete data log likelihood of the d-variate normal distribution is given by

$$L_c(\mathbf{\Psi}; \mathbf{x}, \mathbf{z}) = \sum_{h=1}^{H} \sum_{j=1}^{n^h} \sum_{i=1}^{G} z_{ij}^h ln(\pi_i^h N(\mathbf{x}_j^h; \mathbf{D}_i^{1,h} \boldsymbol{\mu}_i^1 + \mathbf{b}_i^{1,h}, \mathbf{D}_i^{1,h} \Sigma_i^1 \mathbf{D}_i^{1,h})) \tag{14}$$

were the adopted convention is $\mathbf{D}_i^{1,1}$ to be the identity matrix of the appropriate dimension, and $\mathbf{b}_i^{1,1}$ the null vector. The GEM algorithm is outlined as follows

- E-step: Calculate the expected component membership $\hat{z}_{ij}^h(\boldsymbol{\Psi}) = \mathrm{E}[Z_i^h | X^h = \mathbf{x}_j^h, \boldsymbol{\Psi}]$ from the current value of $\boldsymbol{\Psi}$.
- M-step: Substitute the expected component membership, calculated in the E-step, in to $z_{ij}^h$ in equation (12), and then alternatively maximise with respect to the following parameters in $\boldsymbol{\Psi}$; $\{\pi_i^h, \boldsymbol{\mu}_i^1, \Sigma_i^1\}$ and $\{\mathbf{D}_i^{1,h}, \mathbf{b}_i^{1,h}\}$.

Beginning with the reference parameters (Lourme and Biernacki, 2012)

- *Mixing proportions* $\pi_i^1$

  Noting that $\hat{n}_i^h = \sum_{j=1}^{n^h} \hat{z}_{ij}^h$ and $\hat{n}_i = \sum_{h=1}^{H} \hat{n}_i^h$ results in $\pi_i^{1(new)} = \frac{\hat{n}_i^1}{n^1}$ for unrestricted mixing proportions, $\pi_i^{1(new)} = \frac{\hat{n}_i}{n}$ when they are only component dependent and $\pi_i^{1(new)} = \frac{1}{G}$

- *Mean* $\boldsymbol{\mu}_i^1$

$$\boldsymbol{\mu}_k^{1(new)} = \frac{1}{\hat{n}_k} \sum_{h=1}^{H} \sum_{j=1}^{n^h} \hat{z}_{kj}^h (\mathbf{D}_k^{1,h})^{-1} (\mathbf{x}_j^h - \mathbf{b}_k^{1,h})$$

- *Covariance matrix* $\Sigma_i^1$
  For the homoscedastic case the update formula is given by

$$\Sigma^{1(new)} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{G} \sum_{j=1}^{n^h} \hat{z}_{ij}^h ((\mathbf{D}_i^{1,h})^{-1} (\mathbf{x}_j^h - \mathbf{b}_i^{1,h}) - \boldsymbol{\mu}_i^{1(new)})((\mathbf{D}_i^{1,h})^{-1} (\mathbf{x}_j^h - \mathbf{b}_i^{1,h}) - \boldsymbol{\mu}_i^{1(new)})^T$$

  and for the heteroscedastic case

$$\Sigma_k^{1(new)} = \frac{1}{\hat{n}_i} \sum_{h=1}^{H} \sum_{j=1}^{n^h} \hat{z}_{kj}^h ((\mathbf{D}_k^{1,h})^{-1} (\mathbf{x}_j^h - \mathbf{b}_k^{1,h}) - \boldsymbol{\mu}_k^{1(new)})((\mathbf{D}_k^{1,h})^{-1} (\mathbf{x}_j^h - \mathbf{b}_k^{1,h}) - \boldsymbol{\mu}_k^{1(new)})^T$$

The following formulas are used to estimate the link parameter

- *Vectors* $\mathbf{b}_i^{1,h}$
  Noting the empirical mean of component $C_i^h$ as $\bar{\mathbf{x}}_i^h = \frac{1}{\hat{n}_i} \sum_{j=1}^{n^h} \hat{z}_{ij}^h \mathbf{x}_j^h$ the following update formula (unrestricted $\mathbf{b}_i^{1,h}$ with respect to $h$) are given by

$$\mathbf{b}_k^{1,h(new)} = \bar{\mathbf{x}}_k^h - \mathbf{D}_k^{1,h} \boldsymbol{\mu}_k^{1(new)}$$

  and

$$\mathbf{b}^{1,h(new)} = (\sum_{i=1}^{G} \hat{n}_i (\mathbf{D}_i^{1,h} \Sigma_i^{1(new)} \mathbf{D}_i^{1,h})^{-1})(\sum_{i=1}^{G} \hat{n}_i (\mathbf{D}_i^{1,h} \Sigma_i^{1(new)} \mathbf{D}_i^{1,h})^{-1} (\bar{\mathbf{x}}_i^h - \mathbf{D}_i^{1,h} \boldsymbol{\mu}_i^{1(new)}))$$

  for the restricted case.

- *Matrix* $\mathbf{D}_i^{1,h}$
  For the special case where $\mathbf{D}_i^{1,h} = \alpha_i^{1,h}\mathbf{I}$ or $\mathbf{D}_i^{1,h} = \alpha^{1,h}\mathbf{I}$ the respective updates are

$$\alpha_i^{1,h(new)} = \frac{-u_i^h + \sqrt{(u_i^h)^2 - 4d\hat{n_i^h}v_i^h}}{2d\hat{n}_i^h}$$

  or

$$\alpha^{1,h(new)} = \frac{-u^h + \sqrt{(u^h)^2 - 4dn^h v^h}}{2dn^h}$$

  where

$$\begin{aligned} u_i^h &= \sum_{j=1}^{n^h} \hat{z}_{ij}^h (\mathbf{x}_j^h - \mathbf{b}_i^{1,h(new)})^T (\Sigma_i^{1(new)})^{-1} \boldsymbol{\mu}_i^{1(new)} \\ u^h &= \sum_{i=1}^{G} u_i^h \end{aligned}$$

  and

$$\begin{aligned} v_i^h &= \sum_{j=1}^{n^h} \hat{z}_{ij}^h (\mathbf{x}_j^h - \mathbf{b}_i^{1,h(new)})^T (\Sigma_i^{1(new)})^{-1} (\mathbf{x}_j^h - \mathbf{b}_i^{1,h(new)})^T \\ v^h &= \sum_{i=1}^{G} v_i^h \end{aligned}$$

For a general $\mathbf{D}_i^{1,h}$, $\mathbf{D}_i^{1,h(new)}$ can be estimated with any convex optimisation algorithm (Lourme and Biernacki, 2012).

## 2.6 Model Selection

The problem of model selection was address by G. Schwartz (1978) where the Baysian Information Criterion was introduced. The idea behind this is that each model has a score associated with it. This score is given by

$$BIC = -\hat{l}(\Psi, \mathbf{x}) + \frac{\upsilon}{2} log(n)$$

where $\hat{l}(\Psi, \mathbf{x})$ is the maximum likelihood of the model with parameters $\Psi$ of the observed data $\mathbf{x}$. Also, $\upsilon$ are the dimensions of the parameter space and $n$ the total number of observations over all population samples.

## 2.7 Summary

Finite mixture models and Gaussian mixture models for model based clustering were presented. The identifiability problem was formulated and discussed, and several parsimonious models pertaining to the eigendecomposition of the covariance matrix, as given in Banfield and Raftery (1993), were given. Then the data reduction technique factor analysis was introduced. The reasons for the implementation of a factor analytic model are to uncover hidden variables that explain the shape of the data and to model high dimensional data when the usual clustering methods fail due to overparameterisation.

The mathematical definition of single factor analysis was then given, and the identifiability problem was posed and discussed. The idea of factor rotations was discussed, followed by the definition and discussion

of a mixture of factor analysers. The 12 parsimonious models pertaining to subspace modeling, shown in McNicholas and Murphy (2010), was given. Then the method for simultaneous model-based clustering (Lourme and Biernacki, 2012) was examined. The following differences of this model versus independent clustering were given as follows

- One procedure to cluster all samples of all populations as opposed to a different procedure for each sample.
- There are some non-allowed parsimonious models of simultaneous model-based clustering.
- The non-reference samples are not directly clustered in the case of simultaneous model-based clustering.

The notation for the model was then presented, followed by the derivation of the general form of the link function. Some parsimonious models were given and the set of allowable and non-allowable models was given. Following this section was a short account of the definition and applicability of the overlap of clusters (Maitra and Malnykov, 2010). Then the EM algorithms (Dempster, Laird and Rubin, 1977) of model-based Gaussian clustering, clustering by a mixture of factor analysers and simultaneous model-based clustering were given.

# 3 Chapter 3: Simultaneous Model-based Clustering with a Mixture of Factor Analysers

This chapter covers simultaneous clustering with a mixture of factor analysers and describes some parsimonious models associated with it. Section 3.1 derives the defining equations of simultaneous clustering by a mixture of factor analysers. Section 3.2 explores some parsimonious and allowable models. Section 3.3 derives the estimating equation that are used to estimate the link parameters that were derived in section 3.1.

There are cases where it is reasonable to assume there are latent variables in a sample under conditions where multiple, similar samples have been taken in different populations. Clustering a mixture of factor analysers simultaneously may be useful when dealing with data that is assumed to have the same latent variables, yet the loadings on these variables differ from sample to sample. This could also be useful when the clustering of several high-dimensional samples is required. The problem associated with modeling mixtures of factor analysers simultaneously is that if different populations are being considered, it is far more likely that each population will have different latent variables, resulting in poorly clustered data, even if each sample is assumed to have an equal number of factors. As an example, the simultaneous clustering of two or more DNA sample sets may be futile because each sample set may have very different latent variables despite the populations being very similar. It would make more sense to apply the simultaneous method to samples of the same population. The differences between simultaneous model-based clustering by a mixture of normal distributions (shortened to 'simultaneous normal clustering') and that by a mixture of factor analysers (shortened to 'simultaneous FA clustering') can be summarised by 3 main differences. Firstly, the simultaneous normal clustering approach finds a link between the parameters of the different populations that define the normal distribution, namely $\Sigma_i$ and $\mu_i$. The simultaneous FA clustering approach will, instead, finds the link between the parameters of the different populations of the factor analytic model, namely the factor loadings $\mathbf{B}_i$ and the individual specific variances $\psi_i$. Equations 10 and 11 would be nonsensical formulae, and would have to be adapted to calculate the parameters of the FA model for the non-reference samples. This is covered in section 3.1. Secondly, the non-allowable models of the simultaneous FA clustering model are similar to those of simultaneous normal cluster, except it is the combination of the factor analytic parameters of the intrapopulation model with the usual interpopulation model restrictions that can create parsimonious models that do not make sense. For example, for the simultaneous normal clustering model a non-allowed parsimonious model is $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi_k, \Sigma^h\}$. This would have to be translated for the simultaneous FA model to become $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi_k, \mathbf{B}_i^h, \psi_i^h\}$. This is covered in section 3.2. Thirdly, much like the number of parameters of the independent clustering are reduced by implementing the simultaneous normal clustering, so the number of parameters of the simultaneous normal clustering are further reduced by implementing the simultaneous FA clustering. The reduction in the number of parameters from simultaneous normal clustering to simultaneous FA clustering are only from estimating the reference population because the parameters of the model of the non-reference samples are estimated by the parameters of the link function. However, simultaneous model-based clustering with a mixture of factor analysers allows the practitioner to bypass estimating the parameters of the factor model of the subsequent samples, but rather estimate the parameters using the estimates of the link function. A suitable application would be to implement it in an overparameterised model where simultaneous normal clustering would fail.

## 3.1 Mapping Between Multiple Samples

It is important to note the following assumptions before defining the model for simultaneous clustering with FA. Firstly, it is assumed that

1. Each sample of each population can be adequately modeled by an equal number of factors. That is to

say that the latent space is of equal dimension accross samples and therefore equal number of factor loadings.

2. Corresponding factors of each sample has the same real-world meaning. That is that each dimension in the latent space has a physical interpretation, relative to the context of the data, associated with it.

3. The error associated with modeling on a lower dimensional hyper-plane has the same distribution across the models of each sample. Note: The projected points are the ones that are being modeled on the hyperplane. They become the new 'effective' data points that have the distribution $N(0, \mathbf{I}_q)$.

4. The data points that are projected on to the hyper-plane have the same distributions.

Assumptions 3 and 4 are necessary assumptions to ensure that the mapping of parameters from one population to the other has meaning. It would not make sense, for example, to map from a normal distribution to an exponential distribution since the parameters differ. Assumptions 1 and 2 are necessary to ensure that the dimensions of the estimation equations are consistent. The importance of this assumption is demonstrated later.

### 3.1.1 Derivation of Simultaneous Model-based Clustering with a Mixture of Factor Analysers

A mixture of factor analysers for population $h$, $h \in \{1, ..., H\}$, is given by

$$f(\mathbf{x};\mathbf{\Psi}^{\mathrm{h}}) = \sum_{i=1}^{G} \pi_i^h N(\mathbf{x}; \mu_i^h, \mathbf{B}_i^h (\mathbf{B}_i^h)^T + \psi_i^h), \qquad \mathbf{x} \in \mathbb{R}^d$$

The joint moment generating function of $\mathbf{X}_i^{h'}$ is given by (appendix A.3)

$$M_{\mathbf{X}_i^{h'}}(\mathbf{t}) = e^{\mathbf{t}^T (\mathbf{D}_i^{h,h'} \boldsymbol{\mu}_i^h + \mathbf{b}_i^{h,h'}) + \frac{1}{2}\mathbf{t}^T \mathbf{D}_i^{h,h'} \Sigma_i^h (\mathbf{D}_i^{h,h'})^T \mathbf{t}} \tag{15}$$

Applying factor analytic constraint on the covariance matrix gives

$$M_{\mathbf{X}_i^{h'}}(\mathbf{t}) = e^{\mathbf{t}^T (\mathbf{D}_i^{h,h'} \boldsymbol{\mu}_i^h + \mathbf{b}_i^{h,h'}) + \frac{1}{2}\mathbf{t}^T \mathbf{D}_i^{h,h'} (\psi_i^h + (\mathbf{B}_i^h)^T \mathbf{B}_i^h) (\mathbf{D}_i^{h,h'})^T \mathbf{t}} \tag{16}$$

giving,

$$\boldsymbol{\mu}_i^{h'} = \mathbf{D}_i^{h,h'} \boldsymbol{\mu}_i^h + \mathbf{b}_i^{h,h'} \tag{17}$$

$$\mathbf{B}_i^{h'} (\mathbf{B}_i^{h'})^T + \psi_i^{h'} = \mathbf{D}_i^{h,h'} \mathbf{B}_i^h (\mathbf{B}_i^h)^T \mathbf{D}_i^{h,h'} + \mathbf{D}_i^{h,h'} \psi_i^h \mathbf{D}_i^{h,h'} \tag{18}$$

It is reasonable to assume that the $\mathbf{B}_i^{h'}$ of sample $h'$, on the left hand side of equation 18, is not dependent on the noise term $\psi_i^h$ of sample $h$ on the right hand side. Similarly for $\psi_i^{h'}$. To see this the following should be noted; It is assumed in the factor analytic model that the noise term of sample $h$, namely $\epsilon_i^h$, has a distribution independent of $\mathbf{U}_i^h$. Since $\psi_i^h$ is a parameter of $\epsilon_i^h$, and since the factor loadings $\mathbf{B}_i^h$ depend on the factors $\mathbf{U}_i^h$, it is assumed that $\mathbf{B}_i^h$ and $\psi_i^h$ have no relationship. Therefore the mapping solution can be obtained by equating the terms on the left with their respective like terms on the right. However, a unique identifiability problem arises which is addressed in the next section.

### 3.1.2 Identifiability and Proposed Mapping Solution

From equation 18 it can be seen that the following identifiability problem needs to be addressed; Assuming there is an orthogonal $q \times q$ matrix $\mathbf{P}_i^h$ such that $\mathbf{P}_i^h (\mathbf{P}_i^h)^T = \mathbf{I}_q$, the model remains unchanged if we replace

$\mathbf{B}_i^h$ with $\mathbf{B}_i^h\mathbf{P}_i^h$. That is to say equation 18 becomes

$$\mathbf{B}_i^{h'}(\mathbf{B}_i^{h'})^T + \psi_i^{h'} = \mathbf{D}_i^{h,h'}\mathbf{B}_i^h\underbrace{\mathbf{P}_i^h(\mathbf{P}_i^h)^T}_{=\mathbf{I}_q}(\mathbf{B}_i^h)^T\mathbf{D}_i^{h,h'} + \mathbf{D}_i^{h,h'}\psi_i^h\mathbf{D}_i^{h,h'}$$

A proposed mapping solution is as follows;

$$\underset{(p\times q)}{\mathbf{B}_i^{h'}} = \underset{(p\times p)}{\mathbf{D}_i^{h,h'}}\ \underset{(p\times q)}{\mathbf{B}_i^h}\ \underset{(q\times q)}{\mathbf{P}_i^h} \tag{19}$$

$$\underset{(p\times p)}{\psi_i^{h'}} = \underset{(p\times p)}{\mathbf{D}_i^{h,h'}}\ \underset{(p\times p)}{\psi_i^h}\ \underset{(p\times p)}{\mathbf{D}_i^{h,h'}} \tag{20}$$

There is no unique solution to this problem, but this is to be expected since the general factor analytic model is too subjected to a similar identifiability problem. The consistency of the matrix dimensions have been verified. In light of this, the importance of assumption 1 is evident. If population $h'$ is assumed to be modeled by $x$ factors and population $h$ by $y$ factors, where $x \neq y$, then the dimensions in equations 19 and 20 would not agree. Furthermore, it is important to note that the metric for 'similarity of populations' in the case of simultaneous clustering by a mixture of factor analysers is that each population have equal number of similar meaning variables (which is the case for simultaneous model based clustering of Gaussian distributions), but now with the addition that each population be described by the same number of factors. The factors themselves may vary in loadings with each variable but it would be expected that each factor have the same qualitative meaning as its corresponding factor in the previous population. Furthermore, it might be expected that the first criterion in the metric of 'similarity of populations' be relaxed without any loss of meaning. The reason lies in the fact that with the latent variables, the original variables have little meaning in the problem at hand. The model is governed by the latent variables. On the other hand, if the original variables are different in the populations, then we would expect completely different factors and, in general, a different numbers of factors between populations. It is for this reason that both criteria must hold if the simultaneous clustering with a mixture of factor analysers is to hold any meaning. For equation 20 to be feasible the criterion that each population be equal in dimension needs to be met.

## 3.2 Parsimonious Models

Firstly, the most general model is given by $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi_i, \mathbf{B}_i^h, \psi_i^h\}$. The list of non-allowed models of simultaneous FA clustering is similar in principal to those of simultaneous Gaussian clustering. Firstly, it will be nonsensical to have the probability of group membership equal across groups but different across populations. The probability of group membership will have to be 1/K. This can't be different in other populations and therefore any model of the form $\{\pi^h, ..., \pi, ...\}$ is nonsensical. Furthermore, it can be seen from equations 19 and 20 that it would make no sense to have the model $\{..., \mathbf{D}_i^{h,h'}, ..., \mathbf{B}^h, ...\}$. This model would imply that each factor loading matrix and individual specific variance is the same between groups but the transformation from a reference component to the corresponding non-reference component is component-dependent. This model is, also, therefore not allowed. A list of allowed and non-allowed models are given in table 7 which includes all possible combinations of the models from table 5 , as intrapopulation models, with the interpopulation models of the link parameters.

Table 7: *Parsimonious models for simultaneous mixture of factor analysers.*

| | | | $\pi_i$ - $\pi$ | | | |
|---|---|---|---|---|---|---|
| | | | $\mathbf{B}^h$ | | $\mathbf{B}_i^h$ | |
| | | | $\psi^h$ or $\sigma^h\mathbf{I}_p$ | $\psi_i^h$ or $\sigma_i^h\mathbf{I}_p$ | $\psi^h$ | $\psi_i^h$ |
| $\pi(\pi^h)$ | $\mathbf{I},\alpha^{h,h'}\mathbf{I},\mathbf{D}^{h,h'}$ | $\mathbf{0}$ | • (• ) - • (· ) | • (• ) - • (· ) | • (• ) - • (·) | • (• ) - • (·) |
| | | $\mathbf{b}^{h,h'}$ | • (• ) - • (· ) | • (• ) - • (· ) | • (• ) - • (·) | • (• ) - • (·) |
| | | $\mathbf{b}_i^{h,h'}$ | • (• ) - • (· ) | • (• ) - • (· ) | • (• ) - • (·) | • (• ) - • (·) |
| | $\alpha_i^{h,h'}\mathbf{I},\mathbf{D}_i^{h,h'}$ | $\mathbf{0}$ | · (· ) - · (· ) | · (· ) - · (· ) | · (· ) - · (· ) | • (• ) - • (·) |
| | | $\mathbf{b}_i^{h,h'}$ | · (· ) - · (· ) | · (· ) - · (· ) | · (· ) - · (· ) | • (• ) - • (·) |

*Note: "•" represents allowable models and "·" non-allowable models*

Table 7 includes all possible combinations of the models from table 5 , as intrapopulation models, with the interpopulation models. To implement the model $\{\pi^h, \mathbf{D}_i^{h,h'}, \mathbf{b}_i^{h,h'}, \pi_i, \mathbf{B}_i^h, \sigma_i^h\mathbf{I}\}$ equation 20 will change to $\sigma_i^{h'}\mathbf{I} = \sigma_i^h\mathbf{D}_i^{h,h'}\mathbf{D}_i^{h,h'}$. Similarly, for a model such as $\{\pi^h, \mathbf{D}^{h,h'}, \mathbf{b}_i^{h,h'}, \pi_i, \mathbf{B}_i^h, \sigma^h\mathbf{I}\}$ equation 20 would become $\sigma^{h'}\mathbf{I} = \sigma^h\mathbf{D}^{h,h'}\mathbf{D}^{h,h'}$. These two models would account for the case of probabilistic principal components. Furthermore, one might expect the model of equal noise terms across populations to be advisable. This model is of course possible, and plausible, and the result would simply be the omission of equation 20 in the calculations of parameter estimation.

### 3.3 Parameter Estimation

The technique of parameter estimation is adapted from that of Lourme and Biernacki (2012), and utilizes the Alternating Expectation Conditional Maximisation (Meng and VanDyk, 1997). First note that in the first cycle of the AECM algorithm, where the missing data is taken to be $\mathbf{z}$, the likelihood is given by

$$l_1(\mathbf{\Psi};\mathbf{x},\mathbf{z}) = \prod_{h=1}^{H}\prod_{j=1}^{n^h}\prod_{i=1}^{G}[\pi_g f(\mathbf{x}_j^h|\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h})]^{z_{ij}^h} \tag{21}$$

$$= \sum_{h=1}^{H}\sum_{j=1}^{n^h}\sum_{i=1}^{G} z_{ij}^h ln(\pi_i^h N(\mathbf{x}_j^h|\mathbf{D}_i^{1,h}\boldsymbol{\mu}_i^1 + \mathbf{b}_i^{1,h}, \mathbf{D}_i^{1,h}\mathbf{B}_i^1(\mathbf{D}_i^{1,h}\mathbf{B}_i^1)^T + \mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h})) \tag{22}$$

where we have the set of parameters $\mathbf{\Psi} = \{\pi_i^h, \boldsymbol{\mu}_i^1, \mathbf{B}_i^1, \psi_i^1, \mathbf{D}_i^{1,h}, \mathbf{b}_i^{1,h}\}$ for $i \in \{1,...,G\}$, $h \in \{1,...,H\}$. For this step we still have the same estimating equations for mean and probability of component membership;

$$\boldsymbol{\mu}_k^{1(new)} = \frac{1}{\hat{n}_k}\sum_{h=1}^{H}\sum_{j=1}^{n^h}\hat{z}_{kj}^h(\mathbf{D}_k^{1,h})^{-1}(\mathbf{x}_j^h - \mathbf{b}_k^{1,h}) \tag{23}$$

$$\pi_k^{h(new)} = \frac{\hat{n}_k^h}{n^h} \tag{24}$$

where $\hat{n}_k^h = \sum_{j=1}^{n^h}\hat{z}_{kj}^h$, $\hat{n}_k = \sum_{j=1}^{n^h}\sum_{h=1}^{H}\hat{z}_{ij}^h$ , $\hat{n} = \sum_{h=1}^{H}\sum_{j=1}^{n^h}\sum_{i=1}^{G}\hat{z}_{ij}^h$ and where

$$z_{kj}^h = \frac{\pi_k^h N(\mathbf{x}_j^h|\boldsymbol{\mu}_k^1,\mathbf{B}_k^1,\psi_k^1,\mathbf{D}_k^{1,h},\mathbf{b}_k^{1,h})}{\sum_{i=1}^{G}\pi_i^h N(\mathbf{x}_j^h|\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h})} \tag{25}$$

In this cycle $\mathbf{D}_i^{1,h}$and $\mathbf{b}_i^{1,h}$ can also be evaluated in the same way as for the case in Lourme and Biernacki

(2012).

$$\mathbf{b}_k^{1,h(new)} = \frac{1}{\hat{n}_k}\sum_{j=1}^{n^h}\hat{z}_{ikj}^h\mathbf{x}_j^h - \mathbf{D}_k^{1,h}\boldsymbol{\mu}_k^{1(new)} \tag{26}$$

For the second cycle of the AECM algorithm the missing data is taken to be the factors $\mathbf{u}$ and $\mathbf{z}$. This step will be done in more detail due to it's complexity. In this cycle the likelihood equation is given by

$$\begin{aligned} l_2(\boldsymbol{\Psi};\mathbf{x},\mathbf{z}) &= \prod_{h=1}^{H}\prod_{j=1}^{n^h}\prod_{i=1}^{G}[\pi_g f(\mathbf{x}_j^h|\mathbf{u}_j,\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h})]^{z_{ij}^h} &(27)\\ &= \sum_{h=1}^{H}\sum_{j=1}^{n^h}\sum_{i=1}^{G}z_{ij}^h ln(\pi_i^h N(\mathbf{x}_j^h|\mathbf{D}_i^{1,h}\boldsymbol{\mu}_i^1+\mathbf{b}_i^{1,h}+\mathbf{D}_i^{1,h}\mathbf{B}_i^1\mathbf{u}_j^h,\mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h})) &(28)\end{aligned}$$

Taking the expected value of the complete data likelihood function, and after much soul-destroying algebra, we get

$$\begin{aligned} Q(\mathbf{B}_i^1,\psi_i^1) &= \sum_{h=1}^{H}\sum_{j=1}^{n^h}\sum_{i=1}^{G}z_{ij}^h[-\frac{1}{2}log|\mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h}| - \frac{1}{2}tr\{(\mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h})^{-1}(\mathbf{x}_j^h-\mathbf{b}_i^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_i^1)^T(\mathbf{x}_j^h-\mathbf{b}_i^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_i^1)\}\\ &+(\mathbf{x}_j^h-\mathbf{b}_i^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_i^1)^T(\mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h})^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\mathbb{E}[\mathbf{u}_j^h|\mathbf{x}_j^h,\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h}]\\ &-\frac{1}{2}tr\{(\mathbf{B}_i^1)^T\mathbf{D}_i^{1,h}(\mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h})^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\}\mathbb{E}[\mathbf{u}_j^h(\mathbf{u}_j^h)^T|\mathbf{x}_j^h,\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h}]] \end{aligned}$$

We need to evaluate $\mathbb{E}[\mathbf{u}_j^h|\mathbf{x}_j^h,\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h}]$ and $\mathbb{E}[\mathbf{u}_j^h(\mathbf{u}_j^h)^T|\mathbf{x}_j^h,\boldsymbol{\mu}_i^1,\mathbf{B}_i^1,\psi_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h}]$. Following Ghahramani and Hinton (1997) we have

$$\begin{aligned}\mathbb{E}[\mathbf{u}_j^h|\mathbf{x}_j^h,\mathbf{B}_k^1,\psi_k^1,\mathbf{D}_k^{1,h},\mathbf{b}_k^{1,h}] &= (\mathbf{D}_k^{1,h}\mathbf{B}_k^1)^T(\mathbf{D}_k^{1,h}\mathbf{B}_k^1\mathbf{D}_k^{1,h}(\mathbf{B}_k^1)^T+\mathbf{D}_k^{1,h}\psi_k^1\mathbf{D}_k^{1,h})(\mathbf{x}_j^h-\mathbf{b}_k^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_k^1) &(29)\\ &= \beta_k^h(\mathbf{x}_j^h-\mathbf{b}_k^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_k^1) &(30)\end{aligned}$$

and

$$\mathbb{E}[\mathbf{u}_j^h(\mathbf{u}_j^h)^T|\mathbf{x}_j^h,\mathbf{B}_k^1,\psi_k^1,\mathbf{D}_k^{1,h},\mathbf{b}_k^{1,h}] = \mathbf{I}_q-\beta_k^h\mathbf{D}_k^{1,h}\mathbf{B}_k^1+\beta_k^h(\mathbf{x}_j^h-\mathbf{b}_k^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_k^1)(\mathbf{x}_j^h-\mathbf{b}_k^{1,h}-\mathbf{D}_k^{1,h}\boldsymbol{\mu}_k^1)^T(\beta_k^h)^T \tag{31}$$

It is important to note that at this point the parameters $\pi_i^h,\boldsymbol{\mu}_i^1,\mathbf{D}_i^{1,h},\mathbf{b}_i^{1,h}$ have been evaluated in this step. Neatening this equation up we write several elements in shorthand; $\mathbf{D}_i^{1,h}\psi_i^1\mathbf{D}_i^{1,h}=(\psi_i^1)^*,\mathbf{x}_j^h-\mathbf{b}_i^{1,h}=(\mathbf{x}_j^h)^*,\sum_{j=1}^{n^h}\frac{1}{n_k^h}((\mathbf{x}_j^h)^*-\mathbf{D}_i^{1,h}\boldsymbol{\mu}_i^1)^T((\mathbf{x}_j^h)^*-\mathbf{D}_i^{1,h}\boldsymbol{\mu}_i^1)=\mathbf{S}_i^h,\mathbf{I}_q-\beta_i^h\mathbf{D}_i^{1,h}\mathbf{B}_i^1+\beta_i^h\mathbf{S}_i^h(\beta_i^h)^T=\boldsymbol{\Theta}_i^h$

$$\begin{aligned}Q(\mathbf{B}_i^1,\psi_i^1) = \sum_{h=1}^{H}\sum_{i=1}^{G}\hat{n}_i^h[\frac{1}{2}log|((\psi_i^1)^*)^{-1}| - \frac{1}{2}tr\{((\psi_i^1))^{*-1}\mathbf{S}_i^h\}+tr\{((\psi_i^1)^*)^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\beta_i^h\mathbf{S}_i^h\}\\ -\frac{1}{2}tr\{(\mathbf{B}_i^1)^T\mathbf{D}_i^{1,h}((\psi_i^1)^*)^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\boldsymbol{\Theta}_i^h\}] \end{aligned}\tag{32}$$

Maximising $Q(\mathbf{B}_i^1,\psi_i^1)$ with respect to $\mathbf{B}_i^1$ and $\psi_i^1$ we have the following score functions

$$S_1(\mathbf{B}_k^1,(\psi_k^1)^*) = \frac{\partial}{\partial\mathbf{B}_k^1}[\sum_{h=1}^{H}\sum_{i=1}^{G}tr\{((\psi_i^1)^*)^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\beta_i^h\mathbf{S}_i^h\}-\frac{1}{2}tr\{(\mathbf{B}_i^1)^T\mathbf{D}_k^{1,h}((\psi_i^1)^*)^{-1}\mathbf{D}_k^{1,h}\mathbf{B}_k^1\boldsymbol{\Theta}_i^h\}] \tag{33}$$

$$S_2(\mathbf{B}_k^1,(\psi_k^1)^*) = \frac{\partial}{\partial((\psi_k^1)^*)^{-1}}[\sum_{h=1}^{H}\sum_{i=1}^{G}\frac{1}{2}log|((\psi_i^1)^*)^{-1}| - \frac{1}{2}tr\{((\psi_i^1))^{*-1}\mathbf{S}_i^h\}+tr\{((\psi_i^1)^*)^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\beta_i^h\mathbf{S}_i^h\}$$

$$-\frac{1}{2}tr\{(\mathbf{B}_i^1)^T\mathbf{D}_i^{1,h}((\psi_i^1)^*)^{-1}\mathbf{D}_i^{1,h}\mathbf{B}_i^1\mathbf{\Theta}_i^h\}] \tag{34}$$

From these we have

$$S_1(\mathbf{B}_k^1,(\psi_k^1)^*) = \sum_{h=1}^{H} [\mathbf{D}_k^{1,h}(((\psi_k^1)^*)^{-1})^T\mathbf{S}_k^h\hat{\beta_k^h} - ((\psi_k^1)^*)^{-1}\mathbf{D}_k^{1,h}\mathbf{B}_k^1\mathbf{\Theta}_k^h] = 0 \tag{35}$$

$$\Rightarrow \mathbf{B}_k^1 = \sum_{h=1}^{H}\mathbf{S}_k^h\hat{\beta}_k^h(\mathbf{\Theta}_k^h)^{-1} \tag{36}$$

where we have that $\hat{\bullet}$ means the estimate, and where $\mathbf{S}_k^h$ and $\beta_k^h$ are symetric. Furthermore we have

$$S_1(\mathbf{B}_k^1,(\psi_k^1)^*) = \sum_{h=1}^{H}\frac{1}{2}(\psi_k^1)^* - \frac{1}{2}\mathbf{S}_k^h + \mathbf{D}_k^{1,h}\mathbf{B}_k^1\hat{\beta}_k^h\mathbf{S}_k^h - \frac{1}{2}\mathbf{D}_k^{1,h}\mathbf{B}_k^1\mathbf{\Theta}_k^h(\mathbf{B}_k^1)^T\mathbf{D}_k^{1,h} = 0 \tag{37}$$

$$\Rightarrow \sum_{h=1}^{H}(\psi_k^1)^* = \frac{1}{H}\sum_{h=1}^{H}\mathbf{S}_k^h - 2\mathbf{D}_k^{1,h}\mathbf{B}_k^1\hat{\beta}_k^h\mathbf{S}_k^h + \mathbf{D}_k^{1,h}\mathbf{B}_k^1\mathbf{\Theta}_k^h(\mathbf{B}_k^1)^T\mathbf{D}_k^{1,h} \tag{38}$$

$$\Rightarrow \psi_k^1 = \frac{1}{H}\sum_{h=1}^{H}(\mathbf{D}_k^{1,h})^{-1}\mathbf{S}_k^h(\mathbf{D}_k^{1,h})^{-1} - 2\mathbf{B}_k^1\hat{\beta}_k^h\mathbf{S}_k^h(\mathbf{D}_k^{1,h})^{-1} + \mathbf{B}_k^1\mathbf{\Theta}_k^h(\mathbf{B}_k^1)^T \tag{39}$$

### 3.4 Discussion

This section proposed the method of simultaneous clustering with a mixture of factor analysers. Firstly, the differences were between simultaneous normal clustering and simultaneous FA clustering were discussed. These are summarised as follows. Firstly, the interpopulation parameters for the simultaneous FA clustering are used to link the factor analytic parameters of the model for the reference sample to the non-reference sample. Secondly, a difference in allowable and non-allowable models was noted with the allowable models summarised in Table 4. Thirdly, there is a reduction in parameters from the simultaneous normal clustering to simultaneous FA clustering.

The mapping solution for the factor analytic variables between samples was derived and given by equations 19 and 20. Following this, a new metric for 'similarity of population' is defined which must fulfill both of the following criteria

- Each population have equal number of similar meaning variables.
- Each population be described by the same number of factors where each factor is expected to have the same meaning as it's corresponding factor in the reference population.

A set of allowable models was summarised in Table 4 followed by a parameter estimation technique for the model $(\pi,\mathbf{D}^{1,h},\mathbf{b}^{1,h},\mathbf{B}^1,\psi^1)$. Equations 19 and 20 summarise the technique of modeling a mixture of factor analysers simultaneously. In light of the current literature of model-based clustering, equation 19 and 20 are the first of its kind for the case of simultaneous model-based clustering by a mixture of factor analysers.

# 4 Chapter 4: Applications

## 4.1 Simulated Data

The techniques discussed will be first implemented on simulated data. To start off with, simulated data will be clustered using the "mclust" package (Fraley et al, 2012) and the results will be compared with the true parameters. Then, parsimonious Gaussian mixture models will be used to cluster high dimensional data through the package "pgmm" (McNicholas et al, 2011).

### 4.1.1 Model-based Clustering with a Mixture of Normal Distributions

The simulated data were generated from a 5-component bivariate normal mixture model. The true parameter values are given in table 8

Table 8: *True parameters of each of the 5 components of the mixture model*

| Component Number | Mixing proportions | Mean vector | Covariance matrix |
|---|---|---|---|
| 1 | 0.123 | $\begin{pmatrix} 3.811 \\ 5.233 \end{pmatrix}$ | $\begin{pmatrix} 0.259 & 0.211 \\ 0.211 & 0.953 \end{pmatrix}$ |
| 2 | 0.236 | $\begin{pmatrix} 5.323 \\ 8.386 \end{pmatrix}$ | $\begin{pmatrix} 2.446 & -1.418 \\ -1.418 & 1.821 \end{pmatrix}$ |
| 3 | 0.268 | $\begin{pmatrix} 3.168 \\ 8.645 \end{pmatrix}$ | $\begin{pmatrix} 0.483 & -0.423 \\ -0.423 & 0.906 \end{pmatrix}$ |
| 4 | 0.243 | $\begin{pmatrix} 8.213 \\ 7.501 \end{pmatrix}$ | $\begin{pmatrix} 2.110 & -0.043 \\ -0.043 & 0.402 \end{pmatrix}$ |
| 5 | 0.13 | $\begin{pmatrix} 6.869 \\ 9.407 \end{pmatrix}$ | $\begin{pmatrix} 0.148 & -0.061 \\ -0.061 & 0.581 \end{pmatrix}$ |

The data was generated using the "MixSim" function in R (V. Melnykov et al, 2012). First, the 5-component heterogeneous mixture model was estimated with $2.3 < \boldsymbol{\mu}_i < 10$, $i \in (1, ..., 5)$ and $min(\pi_i) = 0.12$. The average overlap $\bar{\omega}_{ij} = 0.09$ with a maximum overlap at $max\{\omega_{ij}\} = 0.3$. From the normal mixture model above, with the parameters of each component given, individual data points were plotted, as shown in Figure 1, using the 'simdataset' function in the MixSim package. A set of 500 points were plotted.

Figure 1: *Scatter plot of two variables from simulated data from 5-component heterogeneous mixture model.*

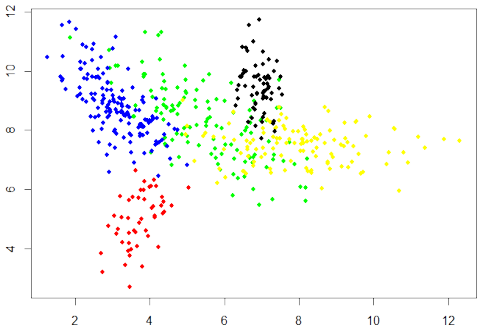
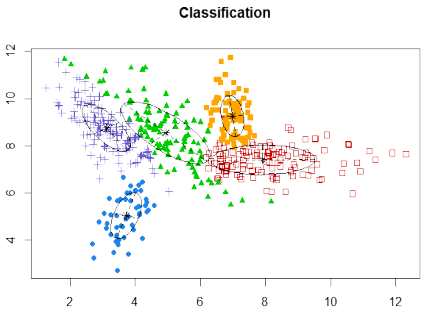


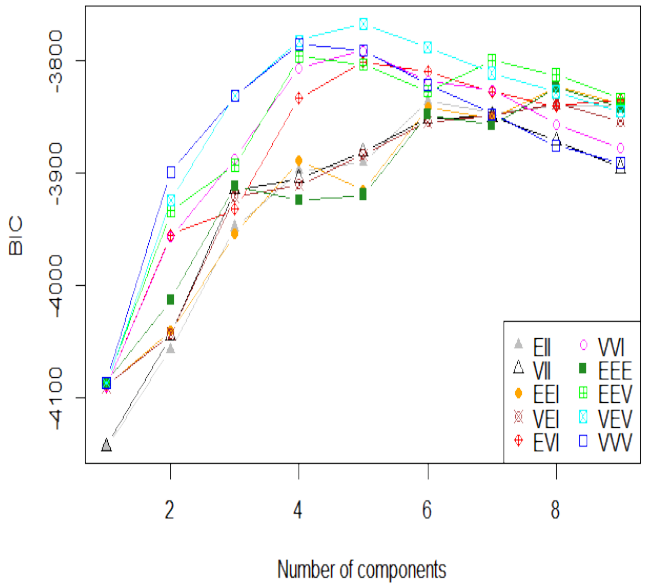

Figure 2: *Clusters found with mclust when classifying the simulated data.*




The goal is to test the robustness of mclust and to uncover the 5 components of the mixture model. The number of components is tested using the BIC. Mclust was performed using the EM algorithm as shown in section 2.5.1. The mclust function was applied to the 500 points of the data matrix and the results are as

follows; The mclust successfully uncovered a 5 component mixture model using the BIC. Figure 2 is a plot of the classifications that mclust uncovered where the clusters where the best model shows the clusters to be all be ellipsoidal, with different sizes and different orientations. Figure 3 is a plot of the BIC values from 1 to 9 components.

Figure 3: *BIC versus number of components for all GMM applied to simulated 5-component data.*




The best model was the 5 component, VEV model with a BIC value of -3766.46. The clustered data is shown in Figure 2. The components uncovered with mclust, and their parameter estimates are given in table 9

Table 9: *The 5 components uncovered with mclust*

| Component Number | Mixing proportions | Mean vector | Covariance matrix |
|---|---|---|---|
| 1 | 0.098 | $\begin{pmatrix} 3.725 \\ 5.022 \end{pmatrix}$ | $\begin{pmatrix} 0.224 & 0.233 \\ 0.233 & 0.952 \end{pmatrix}$ |
| 2 | 0.243 | $\begin{pmatrix} 4.939 \\ 8.547 \end{pmatrix}$ | $\begin{pmatrix} 1.861 & -1.321 \\ -1.321 & 1.738 \end{pmatrix}$ |
| 3 | 0.259 | $\begin{pmatrix} 3.119 \\ 8.742 \end{pmatrix}$ | $\begin{pmatrix} 0.548 & -0.521 \\ -0.521 & 0.994 \end{pmatrix}$ |
| 4 | 0.241 | $\begin{pmatrix} 7.916 \\ 7.392 \end{pmatrix}$ | $\begin{pmatrix} 2.562 & 0.034 \\ 0.034 & 0.393 \end{pmatrix}$ |
| 5 | 0.159 | $\begin{pmatrix} 6.984 \\ 9.257 \end{pmatrix}$ | $\begin{pmatrix} 0.128 & -0.078 \\ -0.078 & 0.767 \end{pmatrix}$ |

Table 9 shows that the estimation of each mixing proportion was somewhat close to the true mixing proportion.

To analyse the error associated with the classification process, it is useful to see the number of misclassified points. Table 10 shows which points where correctly, and incorrectly classified.

Table 10: *Classification Table*

| | | Component Label Using Mclust | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| True Label | 1 | 48 | 0 | 0 | 1 | 0 |
| | 2 | 0 | 23 | 80 | 12 | 10 |
| | 3 | 1 | 0 | 13 | 131 | 0 |
| | 4 | 0 | 93 | 14 | 0 | 10 |
| | 5 | 0 | 2 | 0 | 0 | 62 |

Component 1 generated by the MixSim function corresponds to the component labeled 1 found by clustering the simulated data using Mclust. Similarly, component 2 corresponds to component 3, component 3 to component 4, component 4 to component 2 and component 5 to component 5. Table 10 shows that cluster 1 was accurately exposed with only 1 data point being misclassified to belong to cluster 4. Cluster 2 we widely spread and therefore showed a high misclassification rate with a total of 45 data point misclassified. Cluster 3 was accurately classified with only 15 of the 145 points misclassified. Component 4 was misclassified to have 14 points belong to cluster 3 and 10 points belong to cluster 5. Cluster 5 was very accurate with only 2 of 64 points misclassified. The total number of misclassified points is 86 of a total number of 500 data points. That gives an average of 17.2% misclassification rate. The rate of misclassification could easily be because of the inherent overlap in the system. One would expect the misclassification rate to be close to zero with well separated clusters. Also, the BIC value of the model and the error rate are not necessarily correlated. These are two different measurement. The BIC gives and indication on how the best model to fit the data. This does not necessarily translate in to the lowest rate of misclassification. A table of each model‘s highest BIC value and each misclassification rate are given in Table 11

Table 11: *Top BIC values for each model and their misclassification rates*

| Model | Number of Components | BIC | Misclassification rate |
|---|---|---|---|
| EII | 6 | -3835.538 | 35.6% |
| VII | 7 | -3849.652 | 40.8% |
| EEI | 8 | -3822.322 | 40.6% |
| VEI | 8 | -3837.908 | 47.6% |
| EVI | 5 | -3800.887 | 33.2% |
| VVI | 5 | -3790.539 | 26.4% |
| EEE | 8 | -3823.802 | 42.2% |
| EEV | 4 | -3795.759 | 22.4% |
| VEV | 5 | -3766.460 | 17.2% |
| VVV | 4 | -3785.149 | 13.8% |

In this simulation, the majority of misclassifications were that of component 2. This could largely be due to the amount of overlap that component 2 has with the other clusters. In this example the model VVV did a better job at classifying components 1,3,4 and 5 than the chosen VEV model. Component 2, however, was poorly classified and this model was therefore discarded by merit of the BIC value. It could be worth noting, however, that the VVV model may have some important information when attempting to classify components 1,3,4 and 5 and can disallow certain data points from belonging to the wrong components. Model VEV, however, is the better choice for modeling the population as a whole.

As a second example, the simulated data had 4 components and was 4 dimensional as shown in Figure 4. The average overlap here was given by $\bar{\omega}_{ij} = 0.01$ with a maximum overlap at $max\{\omega_{ij}\} = 0.05$

Figure 4: *Scatter plot of four variables from simulated data from 4-component heterogeneous mixture model.*

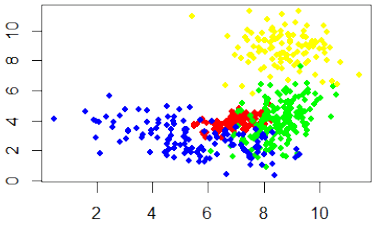



Figure 5: *Clusters found with mclust when classifying the simulated data.*

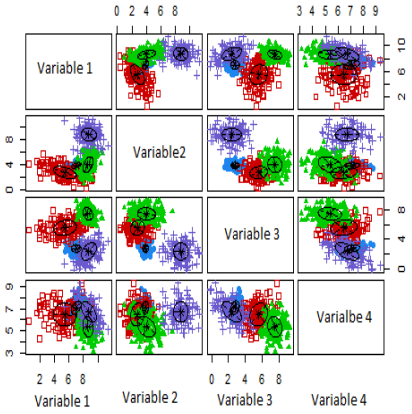



Figure 4 is a 2 dimensional representation of the 4 dimensional data. This is done by projecting the data on to the first two principle components. The BIC values for each model are shown in Figure 5. The VEV shows the highest BIC value with VVV model very close. In fact, the VVV model had the same misclassification rate as the VEV model, however was not chosen because it has more parameters that the VEV model, penalizing it to fall below the VEV model. The component classifications of the VEV model are given in table 12.

Figure 6: *BIC versus number of components for all GMM applied to simulated 5-component data.*

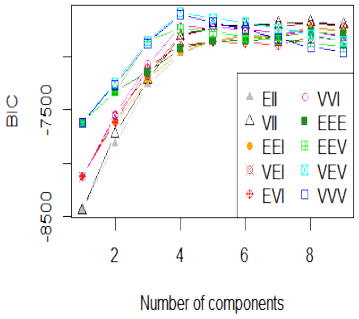



Table 12: *Classification Table*

| | | Clustered Component Label | | | |
|---|---|---|---|---|---|
| | | #1 | #2 | #3 | #4 |
| True Label | #1 | 100 | 1 | 0 | 0 |
| | #2 | 0 | 4 | 162 | 0 |
| | #3 | 0 | 121 | 3 | 0 |
| | #4 | 0 | 0 | 0 | 109 |

This case shows a well classified VEV model with clear component membership and a misclassification rate of 1.6%. The difference in this example was the considerable reduction in the measurement of overlap prescribed to the simulated data. Figure 5 shows 12 plots of each variable against the other along with a

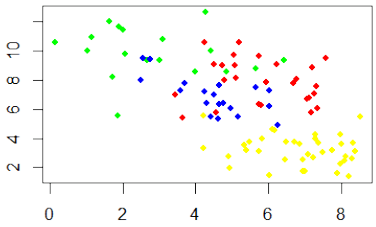
visual representation of the component clusters.

As a final example, the attempt at clustering 100 dimensional data with 50 data points with mclust failed. This model is overparameterised and therefore one should refer to sub-space models that deal with the data that is projected to a lower dimensional subspace.

### 4.1.2 Model-based Clustering with a Mixture of Factor Analysers

It is clear that modeling overparameterised data by a mixture of Gaussian produces inconsistent results. This has to do with with the likelihood increasing without bound (Day, 1969). One solution to modeling high dimensional data is to model the data in a reduced subspace. Table 3 gives a list of parsimonious models pertaining to subspace modeling. These models can be tested on high dimensional data through the 'pgmm' package.
The first set of data points to be generated were those where the number of variables is less than the number of data points. In this data set, the true number of components G = 4, n=100 and the number of variables d=30. There was no prescribed overlap, making the problem, in general, easier to solve because there is no possibility of misclassification of points. The data points are plotted in figure 7 are shown to be overlapping, however this is due to 30 dimensional data being projected on to a plane of two dimensions resulting in seemingly overlapping clusters.

Figure 7: *Scatter plot of thirty variables from simulated data from 4-component heterogeneous mixture model of 100 data points.*




The results of running the pgmmEM algorithm with a k-means start showed a CCU model, with a BIC value of -5883.793, to be favourable with the number of factors q=1 and G=4. The classification table shown in Table 13 shows that no points were misclassified and that the data can be adequately explain by 1 latent variable.

Table 13: *Classification table of the CCUU model with q=1 and G=4*

| | | Clustered Component Label | | | |
|---|---|---|---|---|---|
| | | #1 | #2 | #3 | #4 |
| True Label | #1 | 0 | 26 | 0 | 0 |
| | #2 | 18 | 0 | 0 | 0 |
| | #3 | 0 | 0 | 19 | 0 |
| | #4 | 0 | 0 | 0 | 37 |

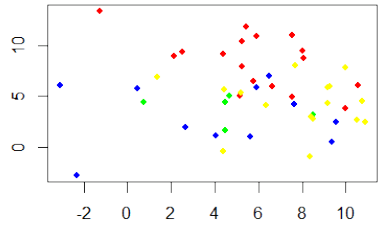
The top BIC values for each model are given in Table 14 where it is clear that each model found a 1 factor model to sufficiently describe the data.

Table 14: *Top BIC values of each model for pgmm algorithm with p=30, n=100*

| Model | Number of components | Number of factors | Maximum BIC value |
|---|---|---|---|
| CCC | 4 | 1 | -6085.692 |
| CCU | 4 | 1 | -5999.303 |
| CUC | 4 | 1 | -5981.171 |
| CUU | 4 | 1 | -6171.298 |
| UCC | 3 | 1 | -6268.803 |
| UCU | 3 | 1 | -6161.482 |
| UUC | 3 | 1 | -6157.334 |
| UUUU | 3 | 1 | -6244.349 |
| CCUU | 4 | 1 | -5883.793 |
| UCUU | 3 | 1 | -6054.349 |
| CUCU | 4 | 1 | -6198.525 |
| UUCU | 3 | 1 | -6336.717 |

The second set of data points are generated by a 4 component mixture model with no overlap. The 50 data points were 100-variate. This example demonstrates the case when there are more dimensions than data points. The data points can not be adequately plotted on a 2 dimensional graph because the plotting function only shows a projection on to the first two vectors that explain most of the variance. For the given example, the number of factors used to explain the variance was 6. The inability to adequately plot this in 2 dimensions is illustrated in figure 6. There is clearly no visible group structure or clusters.

Figure 8: *Scatter plot of high dimensional data plotted on the first two principal axes*




The pgmm method, with k-means start, found a 5 component mixture with 6 factors to explain the variance. The best model chosen was CCUU with a BIC of -5309.225. The classification table is given in Table 15.

Table 15: *Classifications of the CCUU model*

| | | Clustered Component Label | | | | |
|---|---|---|---|---|---|---|
| | | #1 | #2 | #3 | #4 | #5 |
| True Label | #1 | 0 | 0 | 17 | 0 | 0 |
| | #2 | 6 | 0 | 0 | 0 | 0 |
| | #3 | 0 | 6 | 0 | 0 | 5 |
| | #4 | 0 | 0 | 0 | 16 | 0 |

The true clusters labeled 1 and 4 were accurately captured by clusters labeled 3 and 4, respectively, by the pgmm method. Cluster 2 was also found to have no missclassified data points, and was appropriately found to be cluster labeled 1 of the pgmm algorithm. The data points of cluster 3 were scattered amongst clusters labeled 2 and 5. It is here that the misclassifications took place with an almost 50% misclassification rate of this cluster.

Table 16: *Top BIC values for each model*

| Model | Number of components | Number of factors | Maximum BIC value |
|---|---|---|---|
| CCCU | 3 | 1 | -13783.15 |
| CCUU | 3 | 1 | -13995.60 |
| CUCU | 3 | 1 | -13569.28 |
| CUUU | 1 | 3 | -14177.73 |
| UCCU | 1 | 3 | -13961.08 |
| UCUU | 1 | 3 | -14177.73 |
| UUCU | 3 | 1 | -13955.16 |
| UUUU | 1 | 3 | -14177.73 |
| CCUU | 5 | 6 | -5309.225 |
| UCUU | 3 | 1 | -14165.20 |
| CUCU | 1 | 3 | -14177.74 |
| UUCU | 1 | 3 | -14177.74 |

The number of data points was very low (n=50) and was high dimensional (d=100). McLachlan, Peel and Bean (2002) demonstrate the difficulty in dealing with high dimensional data with a low number of observations and comment that it is, in general, a very difficult problem to solve. This problem is the norm when dealing with DNA data. It is common to have high dimensional data with a very low number of observations.

A third example is another illustration of the difficulty in uncovering subgroups in overparameterised data. The simulated data was set to have no overlap, so as to clearly define each subgroup as completely separate from the other subgroups. The number of subgroups was set to 4 with 220 variables and 200 data points. Again, we illustrate an example when $n < d$. The results show a 1 factor, 2 component, UCCU model with the table of misclassifications given in Table 17.

Table 17: *Classifications for n=200 and d=220*

| | | Clustered Component Label | |
|---|---|---|---|
| True Label | | #1 | #2 |
| | #1 | 50 | 0 |
| | #2 | 35 | 0 |
| | #3 | 0 | 43 |
| | #4 | 0 | 72 |

From table 17, it is clear that the pgmmEM algorithm took true clusters labeled 1 and 2, and put them in one cluster labeled cluster #1. Similarly for true clusters 3 and 4.

McNicholas et al. (2010) note that the solution given by the pgmmEM algorithm are sensitive to starting values. The above results were all tested with a k-means start and each only ran once due to the time consuming algorithm being implemented on low-level CPU. This should be retested for several different starting values if one is to attain a well rounded solution.

## 4.2 Real Data

### 4.2.1 Iris data- Clustering with a Mixture of Gaussians

The Iris dataset (Fisher, 1936) was used in the task of testing the classification algorithms. The data set comprised of 150 observation of 4 variables (sepal length, sepal width, petal length, and petal width in cm). There were three different species (Iris Setosa, Iris Versicolour, Iris Virginica) with 50 observation of each. The goal is to cluster the dataset and successfully uncover a three component mixture model which

represents the three different species. The task of clustering was done using the Mclust Package (Fraley et al., 2013).

A Gaussian mixture model was fitted to the Iris dataset and a 2 component 'VEV' model was found to be the best scoring model.

Figure 9: *BIC values of the 10 different models. 2-component VEV model is the best scoring model.*

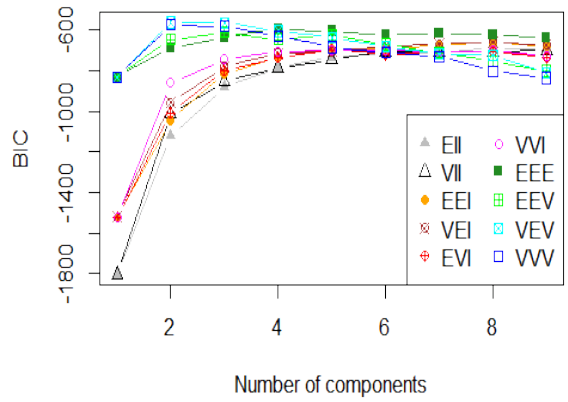



Figure 10: *Clusters found with Mclust. Each figure is the data clustered on the 2D plane with variables on the diagonal.*

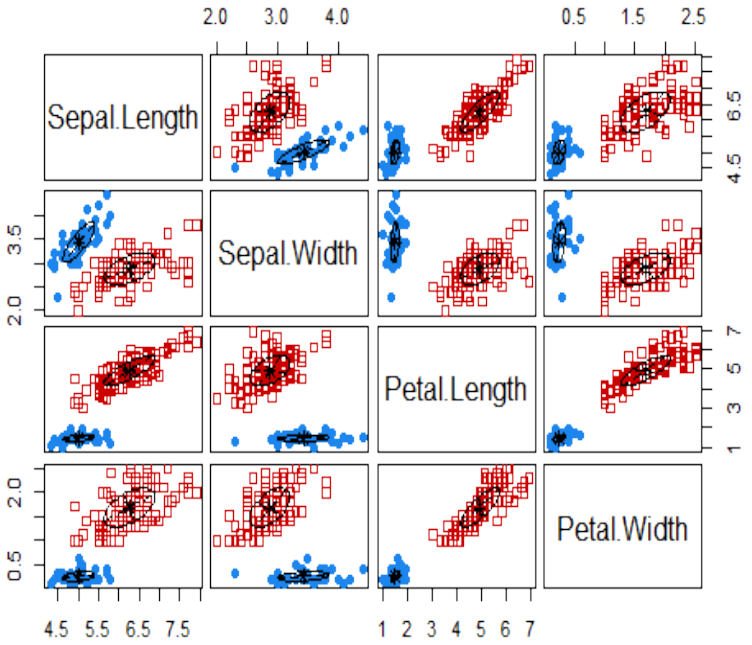



Table 18 shows the classification table and where the misclassifications took place. It is clear that the algorithm combined the species Versicolour and Virginica in to one component.

Table 18: *Classification table of Iris species for the top model (VEV)*

| | | Species | | |
|---|---|---|---|---|
| | | Setosa | Versicolour | Virginica |
| Cluster | 1 | 50 | 0 | 0 |
| | 2 | 0 | 50 | 50 |

The table of the maximum BIC values of each model is given in table 19 along with the misclassification rates.

Table 19: *Classification table of Iris species*

| Model | BIC | Number of components | Misclassification Rate |
|---|---|---|---|
| EII | -686.0967 | 8 | 52.67% |
| VII | -700.022 | 9 | 60.67% |
| EEI | -661.0846 | 8 | 56.67% |
| VEI | -657.2447 | 8 | 50% |
| EVI | -695.6736 | 5 | 30.67% |
| VVI | -696.9024 | 6 | 46% |
| EEE | -591.4097 8 | 4 | 11.33% |
| EEV | -610.0853 | 3 | 2% |
| VEV#1 | **-561.7285** | 2 | 33.33% |
| VEV#2 | -562.5514 | 3 | 3.33% |
| VVV | -574.017 | 2 | 33.33% |

The model VEV#2 was included because this was the best scoring 3 component model. It is clear that the BIC values of the two VEV models are very close but the VEV#2 model describes the data with more accuracy. The VEV#2 model misclassified 5 Versicolour observations as Virginica as shown in table 19. In particular the observations labeled 69, 71, 73, 78, 84 were misclassified. The centers for the VEV#2 model are given in table 20. Upon inspection of the misclassified observations, it is clear that the 5 Versicolour species are very similar to the mean values of the Virginica species and would therefore be expected to be misclassified. Figure 11 shows the clusters generated by the VEV#2 model.

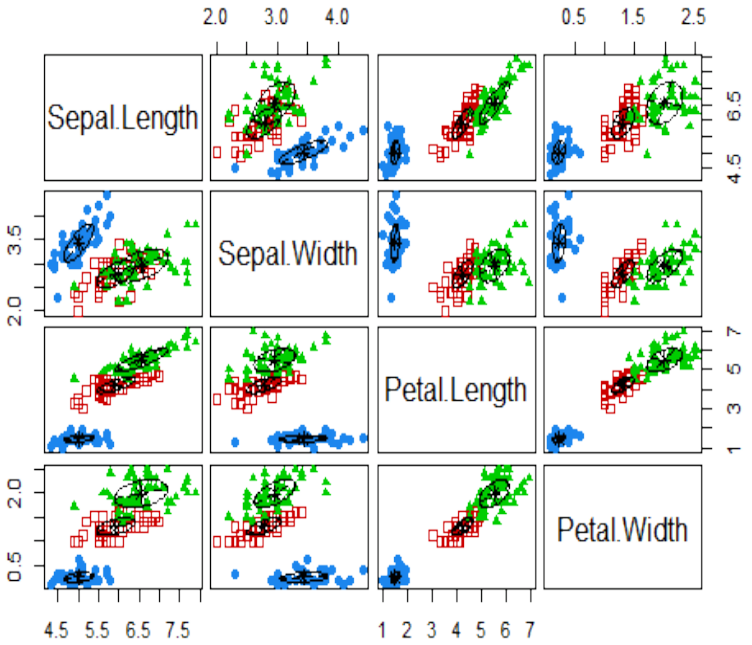
Table 20: *Centers of each Cluster of the VEV#2 model*

| | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Sepal.Length | 5.006 | 5.915 | 6.547 |
| Sepal.Width | 3.428 | 2.778 | 2.949 |
| Petal.Length | 1.462 | 4.204 | 5.482 |
| Petal.Width | 0.246 | 1.299 | 1.985 |

Table 21: *Classification table of Iris species for the VEV#2 model*

| | | Species | | |
|---|---|---|---|---|
| | | Setosa | Versicolour | Virginica |
| Cluster | 1 | 50 | 0 | 0 |
| | 2 | 0 | 45 | 0 |
| | 3 | 0 | 5 | 50 |

Figure 11: *Clusters found with VEV#2 model*




While the VEV#1 model was the preferred model according to the BIC, the VEV#2 model definitely gave a better classification. The number of parameters to estimate the VEV#1 model was 26 versus the 38 parameters for the VEV#2. This could demonstrate the reasons for the lower BIC score of the VEV#2 which penalises the model for each component.

### 4.2.2 Colon Cancer - Mixture of Factor Analysers

The colon cancer data set is given by Alon et al (1999). The data set comprises of over 7500 different gene expressions from 62 tissue samples of which 40 are from tumours and 22 are healthy. Alon et al (1999) retain 2000 genes with the highest minimal intensity. Our data matrix consist of 62 rows where each row is an observation and 2000 columns. Clustering this with a mixture of normal distributions would not be a good idea because the noise of the data set, caused by high dimensionality, would be far too high with only a very small number of samples (McLachlan et al, 2001). To overcome this issue, a mixture of factor analysers will be fitted to the data that will hopefully uncover the relevant subgroups. While this method has been proposed in the past with favourable results, it is still subject to problems pertaining to the curse of dimensionality. As a supplementary technique to reducing the dimensions of the data to a subspace of factors, feature selection is a popular technique for data reduction. Feature selection methods seek out only the relevant features that explain the relevant cluster compositions pertaining to the desired data discrimination. Although feature selection is a relatively unexplored field in the unsupervised context (Villar et al, 2009), there are at least two papers dealing with feature selection of genetic data, namely; McLachlan et al., (2001) and Alladi et al., (2008). Feature selection is regrettably not covered in this work, but the idea of McLachlan et al., (2001), is to test the difference in likelihoods of the one component model versus two component model. If the likelihood of the two-component model is sufficiently larger than that of a one-component model for each feature considered individually, and that each cluster of the two-component model is sufficiently big, then the feature is retained. At this point it is important to note that the number of groups of the data set are not known. It seems contradictory since when performing feature selection we assume the number of groups

to be known. The aim of this exercise is merely to test the robustness of the clustering algorithms, which performed hopelessly without the feature selection. This scenario is limited for this reason but still provides insight in to the particular algorithm.

This section will cluster the colon cancer data set by a mixture of factor analysers using the 'pgmm' package in R. The ideal clustering partition is a two component cluster, where one component indicates healthy tissue and the other unhealthy tissue. When the attempt was made to cluster the gene expression data without feature selection, testing between 60 and 70 factors, the algorithm ran for approximately 98 hours (about 4 days) without an optimal solution. Perhaps this is due to there being more latent variables than samples themselves. When the attempt was made to shorten the intervals of the search range of factors, the maximum number of factors was chosen, by the algorithm, 6 consecutive times with each run taking between 24 and 48 hours. The computer on which the algorithms were run had a i5 3.1GHz CPU and 8 GB RAM. Due to the incredibly long run-times, feature selection was a necessary implementation which was undertaken by the FSelector package (Romanski, 2013).

Before the data is to be mined, it must be suitably prepared for the feature selection process, and then further prepared for the process of modeling with a mixture of factor analysers. The first part of the preparation is to assign each of the 62 samples a "diagnosis" variable that indicates if each sample is one of a tumour tissue or normal tissue. This is so the feature selection can suitably estimate the weights of each variable of a predictive formula that estimates the diagnosis of each sample based on each of the 2000 genes. The features of the retained variables are then gathered together in a new data matrix. After this step, the remaining dataset was standardised to model with a mixture of factor analysers. The reasons for doing this are described in Chapter 2. The FSelector Package, using the 'information.gain' method and selecting the best subset of features using the method 'cutoff.biggest.diff', reduced the dimensions of the dataset from 2000 to 135. This process took approximately 109 seconds. After this, the data was standarised and the 'pgmmEM' algorithm was used to estimate the parameters of the factor analytic model. The algorithm tested between 1 and 10 factors with 2 different k-means starts. This took another 1440.18 seconds (approximately 24 minutes).

The pgmmEM algorithm found that a 8 factor, 3 component, UUCC model was best. Although the cluster did not retrieve the desired discrimination, table 22 shows that there is certainly some predictive power in the model

Table 22: *Classification table of the 3-component model of cancerous tissue*

| | Cluster number | | |
|---|---|---|---|
| | #1 | #2 | #3 |
| Normal | 16 | 0 | 6 |
| Tumour | 3 | 9 | 28 |

From table 22, one would postulate that cluster #1 refers to the cluster of normal tissues. However, 6 normal tissues were misdiagnosed as being cancerous. Similarly, cluster #3 refers to the cancerous tissue where 3 were misdiagnosed as healthy and 9 were unclassified. The problem with the data set is that 22 (tissues labeled 1-22) tissue samples were observed using a poly detector, while the remaining observations were observed using total extraction of RNA (labeled 23-62). McLachlan fitted a 2 component normal mixture model to the reduced data set he obtained from the EMMIX-GENE algorithm and found that the clusters almost entirely represented the genes observed with each method. There is no doubt that the change in the style of observation could be a contributing factor to misclassifications of the observations, but the clustering the 135 pre-selected genes with a mixture of factor analysers has given someone consistent results.

When a two component model was fitted the model was tested for 1 to 20 factors. This took approximately 57 minutes to run. A 19 factor model was found to be favourable. Table 23 shows the misclassifications of

the model

Table 23: *Classification table of the 2-component model of cancerous tissue*

| | Cluster number | |
|---|---|---|
| | #1 | #2 |
| Normal | 21 | 1 |
| Tumour | 20 | 20 |

The model can clearly find half of the cancerous tissue as cluster #2, however cluster #1 contained the same amount of normal tissue as it did tumour tissues. The clusters in table 23 also do not correspond to the different protocols of observation as found by McLachlan et al. (2001). Fitting a two component model with the selected features does not partition the data favourably and one would therefore refer to the 3-component model as the best estimate of the representation of the data.

# 5 Chapter 5: Discussion

## 5.1 Summary

The claim of this dissertation is restated.

> Simultaneous model-based clustering (Lourme and Biernacki, 2012) which allows one to model multiple similar population samples simultaneously by the use of a sufficiently simple inter-population link function, can be extended from a mixture of Gaussian distributions to a Gaussian mixture of factor analysers to account for overparameterised models.

This claim leads to a set of the following of sub-claims:

1. Simultaneous model-based clustering by a mixture of factor is desirable and relevant.
2. There is a link function that links two similar populations that are modeled by a mixture of factor analysers, and the parameters of this mixture model can be estimated through the link function.
3. There is a suitable iterative procedure to estimate the parameters of the model.

The major claim of this work was suitably defended in Chapter 3 by showing how simultaneous model-based clustering can be extended to simultaneous model-based clustering with a mixture of factor analysers. Therein the model was formulated as an extension of the work of Lourme and Biernacki (2012). A parameter estimation scheme was introducted, which should be seen as supplementary to those shown in Lourme and Biernacki (2012), save for the estimation of the covariance matrices being replace by that of the factor loadings and noise parameters.

Sub-claim 1 was defended by showing the importance and natural extension of the current literature to simultaneous model-based clustering with a mixture of factor analysers by demonstrating how the mathematical literature has lead up to this point. The literature on the applications of model-based clustering details the usefulness of the method in the context of modeling multiple samples of gene expression data. However, the usefulness was not able to be appropriately tested. Sub-claim 2 and sub-claim 3 was defended by deriving equations 19 and 20 in Chapter 3 and giving an adapted estimation procedure to estimate the link functions.

Some of the mathematical data mining techniques introduced in Chapter 2 were tested in Chapter 4 which was tested by clustering generated data, as well as real data. The Iris dataset (Fisher, 1936) was found to be best represented by a 2-component model according to the score of the BIC values of each model. However, when a 3-component model was fitted, the misclassification rate dropped by a factor of 10 with little practical penalties caused by over fitting. The colon caner dataset (Alon et al., 1999) was classified using a mixture of factor analysers of a pre-selected subset of genes. The feature selection process found 135 genes to be relevant in deciding the outcome of a diagnosis. A model was then fitted using the 'pgmmEM' function. The best scoring model was the UUCC, 8 factors, and 3 component model. The cluster #2 was labeled as unclassified observation despite the overall model giving a somewhat predictive classifier.

## 5.2 Strengths

The strengths of the work presented in Chapter 3 arise in the usefulness of simultaneous model-based clustering with a mixture of factor analysers pertaining to multiple samples of high dimensional data, such as gene expression data. The model was rigourously formulated, and summarised by equations 19 and 20. At least three strengths are listed;

- Simultaneous model-based clustering with a mixture of factor analysers is used when clustering multiple samples of high dimensional data simultaneously. It could potentially provide a computationally

cheaper way to model multiple high-dimensional samples. However, the new estimation procedure is computationally expensive which may lead to a more expensive algorithm.

- A further reduction in the number of parameters in the factor model of the non-reference samples is notable. The parameter difference is clear when considering that the parameters of factor model of a non-reference sample need not be estimated directly, nor does the number of factors. The number of factors are assumed equal, and the parameter estimation of the factor model is undertaken by estimating the parameters of the link function.
- The link parameters are sufficiently simple and can be estimated using the AECM algorithm.

## 5.3 Weaknesses

There are at least five weaknesses of this analysis.

- Simultaneous model-based clustering with a mixture of factor analysers was not tested in this analysis and therefore the model's usefulness can not be verified.
- Equation 20 may be redundant in the analysis. Assuming equal noise variables across samples is an intuitive assumption. This model was discussed as a parsimonious model in section 3.2 but was not shown to be true or false.
- The data that the model is designed to mine is inherently noisy and computationally expensive. Gene expression data is well known to have many more variables than observations. While feature selection and factor analysis are techniques used to manage the high dimensionality, there will always be a significant error in the estimation procedure which is only perpetuated when using those results to estimate the parameters of the model of another sample.
- The assumption that each sample of each population have the same number of factors is one of convenience. This assumption could pose to be a major flaw in the model if it is found to be a false assumption.

## 5.4 Future Possibilities

Future analyses should entail verifying the model presented as equations 19 and 20 by testing on it on two or more samples and comparing the results with that of independent model-based clustering with a mixture of factor analysers. Several parsimonious models will need to be tested with particular interest in a mixture of principal components.

Further analysis should be done on relaxing equation 20 as a condition of the model. The estimation scheme can be generalised to encompass all inter- and intra-population models as opposed to the two models considered in section 3.3.1. Furthermore, a rigorous analysis of unsupervised feature selection can be supplemented with the proposed model to further reduce high dimensional data. However, the aforementioned analyses would have to be explored in conjunction with the practical issues of high dimensional data (noise, expensive, etc).

A final analysis should be done on the assumption that each population can be modeled with equal number of factors (assumption 1 of section 3.1). If this assumption is found to be false, the analysis of simultaneous model-based clustering with a mixture of factor analysers could prove to be a difficult model to present, if possible at all.

# References


[1] H. Abdi. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences*, 3(1):1–8, 2003.

[2] M. Aitkin and D. B. Rubin. Estimation and Hypothesis Testing in Finite Mixture Models. *Journal of the Royal Statistical Society*, 47(1):67–75, 1985.

[3] S. M. Alladi, P. S. Santosh, V. Ravi, and U. S. Murthy. Colon cancer prediction with genetic profiles using intelligent techniques. *Biomedical Informatics Publishing Group*, 3(3):130–133, 2008.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–50, 1999.

[5] J. Baek, G. J. Mclachlan, and L. K. Flack. Mixtures of factor analyzers with common factor loadings : applications to the clustering and visualisation of high-dimensional data. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1298 – 1309, 2008.

[6] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *AAAI Press Technical Reports*, 2:28–36, 1994.

[7] J. D. Banfield and A. E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *International Biometric Society*, 49(3):803–821, 1993.

[8] F. Beninel, C. Biernacki, C. Bouveyron, Jacques J., and Lourme A. *Parametric link models for knowledge transfer in statistical learning.* chez Nova Publishers, knowledge edition, 2012.

[9] C. Biernacki, F. Beninel, and V. Bretagnolle. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2):387–97, June 2002.

[10] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

[11] C. Biernacki and A. Lourme. Simultaneous t-Model-Based Clustering for Time Dependent Data : Application to a Study of the Financial Health of Corporations. *CS-BIGS*, 4(2):73–82, 2011.

[12] M. Blume. Expectation maximization: a gentle introduction, 2002.

[13] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, December 2012.

[14] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[15] G. M. Church, Y. Gao, and S. Kosuri. Next-generation digital information storage in DNA. *Science (New York, N.Y.)*, 337(6102):1628, September 2012.

[16] N. E. Day. Estimating the Components of a Mixture of Normal Distributions. *Biometrika*, 56(3):463–474, 1969.

[17] B. de Meyer, B. Roynette, P. Vallois, and M. Yor. On independent times and positions for Brownian motions. *Revista Matemática Iberoamericana*, 18(3):541–586, 2002.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[19] A. D'Souza. Derivation of Maximum Likelihood Factor Analysis using EM. pages 1–7, 2000.

[20] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[21] C. Fraley. Algorithms for Model-Based Gaussian Hierarchical Clustering. *SIAM Journal on Scientific Computing*, 20:270–281, 1998.

[22] C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578–588, 1998.

[23] C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, June 2002.

[24] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca. mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation, 2012.

[25] Z. Ghahramani and G. E. Hinton. EM algorithm for Mixture of Factor Analyzers. *Technical Report*, pages 1–8, 1997.

[26] G. E. Hinton, M. Revow, and P. Dayan. Recognizing Handwritten Digits Using Mixtures of Linear Models. *Advances in Neural Information Processing Systems 7*, pages 1015–1022, 1995.

[27] P. J. Huber. Projection Pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[28] M. B. Kursa and W. R. Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistics*, 36(11):1–13, 2010.

[29] G. Lee and C. Scott. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829, 2010.

[30] A. Lourme. *Applied Contribution to the Classification by Mixture Models and Mathematics Simultaneous Classification of Samples of Multiple Origins*. Doctor of philosophy, Universite Lille, 2011.

[31] A. Lourme and C. Biernacki. Simultaneous Gaussian model-based clustering for samples of multiple origins. *Computational Statistics*, 28(1):371–391, February 2012.

[32] J. MacQueen. Some methods for classification and analysis of multivariate observation. *Berkeley Symposium on Mathematical Statistics and Probability*, 233(233):281–297, 1967.

[33] R. Maitra and V. Melnykov. Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics*, 19(2):354–376, January 2010.

[34] F. H. C. Marriott. Separating Mixtures of Normal Distributions. *Biometrics*, 31(3):767–769, 1975.

[35] G. J. McLachlan, R. W. Bean, and D. Peel. Clustering of microarray expression data. *Oxford Journals*, 18(3):413–422, 2001.

[36] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions Second Edition*. 2008.

[37] G.J. McLachlan, D. Peel, and R.W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4):379–388, January 2002.

[38] P. D. McNicholas. Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, 140(5):1175–1181, May 2009.

[39] P. D. McNicholas. On Model-Based Clustering, Classification, and Discriminant Analysis. *Journal of Iranian Statistical Society*, 10(2):181–199, 2011.

[40] P. D. McNicholas, K. R. Jampani, A. F. McDaid, T. B. Murphy, and L. Banks. *pgmm: Parsimonious Gaussian Mixture Models*, 2011. R package version 1.0.

[41] P. D. McNicholas and T. B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, April 2008.

[42] P. D. McNicholas and T. B. Murphy. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics (Oxford, England)*, 26(21):2705–2712, 2010.

[43] V. Melnykov, W. Chen, and R. Maitra. MixSim: Simulating Data to Study Performance of Clustering Algorithms. *Journal of Statistical Software*, 51(12.):1–25, 2012.

[44] V. Melnykov and R. Maitra. Finite Mixture Models and Model-Based Clustering. *Statistical surveys*, 0(0000):0–36, 2010.

[45] X. Meng and D. Van Dyk. The EM Algorithm - an Old Folk-song Sung to a Fast New Tune. *Royal Statistical Society*, 59(3):511–567.

[46] K. P. Murphy. Mixture models, 2006.

[47] I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC bioinformatics BioMed Central*, 8(61), January 2007.

[48] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[49] A. J Scott and M. J. Symons. Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, 27(2):387–397, 1971.

[50] C. Spearman. "General Intelligence", Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[51] M. J. Symons. Clustering Criteria and Multivariate Normal Mixtures. *Biometrics*, 37(1):35–43, 1981.

[52] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1998.

[53] J. A. Van Franeker and C. J . F. Ter Braak. A Generalized Discriminant for Sexing Fulmarine Petrels from External Measurements. *The Auk*, 110(3):492–502, 1993.

[54] M. Varjokallio and M. Kurimo. Comparison of Subspace Methods for Gaussian Mixture Models in Speech Recognition. *Proc. Interspeech*, (3):2121–2124, 2007.

[55] J. K. Vermunt and J. Magidson. *Hierarchical Mixture Models for Nested Data Structures*. Springer Berlin Heidelberg, part ii edition, 2005.

[56] J. R. Villar, M. R. Suarez, J. Sedano, and F. Mateos. Unsupervised Feature Selection in high dimensional spaces and uncertainty. *Artificial Intelligence*, 5572(4):565–572, 2009.

[57] J. H. Ward. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[58] J. H. Wolfe. Pattern Clustering By Multivariate Mixture Analysis. *Multivariate Behavioral Research*, 5(3):329–350, April 1970.

[59] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics (Oxford, England)*, 17(10):977–87, October 2001.

# Appendix A

## A.1 Linear Map Between Two Normal Distributions

"Let $Y \sim N(0,1)$ and $Y \sim \phi(Y)$, where $\phi\ \mathbb{R} \to \mathbb{R}$ and continuously differentiable. Therefore $\phi(y) = \pm y$. Similarly, the result can be extended if $Z \sim \phi(X)$ where $X \sim N(\mu_X, \sigma_X^2)$ and $Z \sim N(\mu_Z, \sigma_Z^2)$. In this case the linear relationship can be written as$\phi(x) = ax + b$.

Proof: Suppose $\phi$ is not monotone. Then there would exist a point $a \in \mathbb{R}$ such that $\phi'(a) = 0$ and so $\phi(Y)$ has an infinite density at $\phi(a)$. Denote the cumulative distribution by $F(\phi(a))$ with $F'(\phi(a)) = \phi'(a)f(\phi(a)) = 0$ for a finite $f(\phi(a))$. Now suppose that $\phi(a)$ is increasing. Then $F(a) = Pr[Y < a] = Pr[\phi(Y) < \phi(a)] = F(\phi(a))$. Therefore $\phi(a) = a$ since the cumulative function is unique. Conclude now by assuming $\phi$ is now decreasing." (Biernacki et al, 2003)

## A.2 Likelihoods, Parameters and Parameter estimation of the Factor analytic model

It will be shown how the likelihood of the conditional and unconditional factor analytic model is found, as well as the estimated means and covariance structures of each case. Note the factor analytic model for data vector $\mathbf{Y}_j$

$$\mathbf{Y}_j - \boldsymbol{\mu}_j = \mathbf{u}_{ij}\mathbf{B}_i + \epsilon_{ij}$$

### Conditional Expectation

In this case it is assumed that the factor scores, $\mathbf{u} \in \mathbb{R}^q$, are known

#### Mean and covariance

$$E[\mathbf{x}|\mathbf{u}] = E[\mathbf{u}B + \epsilon|\mathbf{u}] = \mathbf{u}B$$

and

$$Cov[\mathbf{x}|\mathbf{u}] = E[(\mathbf{x} - \mathbf{u}B)(\mathbf{x} - \mathbf{u}B)^T|\mathbf{u}] = E[\epsilon\epsilon^T|\mathbf{u}] = \psi$$

This gives us the complete factor analytic model $X \sim N(\mathbf{u}B, \psi)$.

### Unconditional Expectation

In this situation, assume the factor scores are unknown and treat them as missing data.

$$\begin{aligned} E[\tfrac{1}{n}Y^TY] &= \frac{1}{n}E[(\epsilon^T + B^TU^T)(UB + \epsilon)] \\ &= \frac{1}{n}(E[\epsilon^T\epsilon] + B^TE[U^T\epsilon] + E[\epsilon^TU]B + B^TE[U^TU]B) \\ &= \psi + 0 + 0 + \frac{1}{n}B^TnIB \\ &= \psi + B^TB \end{aligned}$$

## A.3 Joint Moment Generating Function of the Multivariate Normal Distribution

Let $\mathbf{X} = N_p(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{t} = (t_1, t_t, ...t_p)$. Since $\Sigma$ is positive-semi definite the unique Cholesky decomposition $\Sigma = PP^T$ is given, where $P$ is a lower triangular matrix with real positive diagonal elements. Then there

exists a linear map such that $\mathbf{Y} = P^{-1}(\mathbf{X} - \boldsymbol{\mu})$ where $\mathbf{Y} \sim N_p(\mathbf{0}, I)$. The proof for this involves a direct, analytical computation of the mean and covariance of $\mathbf{Y}$ using the unique Cholesky decomposition (the proof will be omitted in this text).

The joint moment generating function of $\mathbf{Y}$ is then given by

$$\begin{aligned} M_{\mathbf{Y}}(\mathbf{t}) &= E[e^{Y_1 t_1 + Y_2 t_2 + \ldots + Y_p t_p}] = E[e^{\mathbf{t}^T \mathbf{Y}}] \\ &= \prod_{n=1}^{p} E[e^{Y_n t_n}] \\ &= \prod_{n=1}^{p} e^{\frac{1}{2} t_n^2} \end{aligned}$$

given in condensed form

$$M_{\mathbf{Y}}(\mathbf{t}) = e^{\frac{1}{2}\mathbf{t}^T \mathbf{t}}$$

To find the moment generating function of $\mathbf{X}$ the above mapping is used

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= E[e^{\mathbf{t}^T \mathbf{X}}] \\ &= E[e^{\mathbf{t}^T (\boldsymbol{\mu} + P\mathbf{Y})}] \\ &= e^{\mathbf{t}^T \boldsymbol{\mu}} E[e^{(P^T \mathbf{t})^T \mathbf{Y}}] \\ &= e^{\mathbf{t}^T \boldsymbol{\mu}} e^{\frac{1}{2}\mathbf{t}^T P P^T \mathbf{t}} \end{aligned}$$

arriving to

$$M_{\mathbf{X}}(\mathbf{t}) = e^{\mathbf{t}^T \boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T \Sigma \mathbf{t}}$$