Modelling Severe Asthma Variation

Thesis submitted for the degree of Doctor of Philosophy

At the University of Leicester

By Christopher James Newby
Department of Health Sciences

October 2011

Abstract

Asthma is a heterogeneity disease that is mostly managed successfully using bronchodilators and anti-inflammatory drugs. Around 10%-15% of asthmatics however have difficult or severe asthma which is less responsive to treatments. Asthma and in particular severe asthma are now thought of a description of symptoms which may contain possible sub-groups with possible different pathologies which could be useful for targeting different drugs for different sub-groups. However little statistical work has been carried out to determine these sub-phenotypes.

Studies have been carried out to partition severe asthma variables in to a number of sub-groups but the algorithms used in these studies are not based on statistical inference and it is difficult to select the number of best fitting sub-groups using such methods. It is also unclear where the clusters or sub-groups returned are actual sub-groups or reflect a bigger non-normal distribution. In the thesis we have developed a statistical model that combines factor analysis, a method used to obtain independent factors to describe processes allowing for variation over variables, and infinite mixture modelling, a process that involves determining the most probable number of mixtures or clusters thus allowing for variation over individuals. This model created is a Dirichlet process normal mixture latent variable model DPNMLVN and it is capable of determining the correct number of mixtures over each factor. The model was tested with simulations and used to analysis two severe asthma datasets and a cancer clinical trial. Sub-groups were found that reflect a high Eosinophilic group and an average eosinophilic group, a late onset older non atopic group and a highly atopic younger early onset group. In the clinical trial data 3 distinct mixtures were found relating to existing biomarkers not used in the mixture analysis.

Acknowledgements

I would like to thank many people who have helped me along the way as this thesis could not have been possible without their help and encouragement.

I wish to thank my primary supervisor Prof John Thompson for moulding my formless enthusiasm and energy for the project in to something logical, presentable and readable and for guiding me through the 4 year project from inception to completion and beyond.

Thanks go to Astra Zeneca and in particular Dr Athula Hearth for sponsoring my PhD and to Dr Ruth Green and Prof John Thompson for setting up my PhD.

Thanks also go to my second supervisor, Prof Chris Brightling for taking over as second supervisor from Dr Ruth Green in my first year, supporting me through my project and beyond and for helping out with all the clinical issues involved in my work.

All involved in the thesis were incredible patience and kind throughout my PhD experience.

A big thank you to all who supported me and encouraged me especially family and friends and all who work for the University of Leicester especially people in Health Sciences, Infection, Immunity and Inflammation, Student Support and Development and the University Bookshop. This was greatly appreciated.

Contents

Chapter 1. Introduction

1.1 Chapter summary

This chapter gives a brief outline of the aims of the thesis along with a brief account of what is already known on the topic of both severe asthma variation and factor/cluster/mixture modelling methodology for the purpose of analysing severe asthma variation. The chapter then introduces the key concepts and ideas for the thesis in order to gain a base knowledge to understanding the rest of the thesis and further chapters. The overview of the thesis structure is stated in order to act as a guide to reading and to find specific chapters in the thesis.

1.2 Aim of the thesis

Briefly the aim of the thesis is to explore and statistically model the variation between patients with severe asthma in order to determine if the variation found in clinic (see chapter 2) is due to individual severity or distinct clusters or sub-groups. Although a simple biological question the statistical methodology needed to answer this question is both complicated and difficult to compute. The statistical way the thesis addressed this question is by applying semi-parametric latent variable models to explore the patterns of variation in two multivariate severe asthma clinical datasets.

A latent variable is a hypothesised variable that cannot be measured but can be quantified on an arbitrary scale using other outcome variables that are correlated to the latent variable (see chapter 4). Using this concept the latent variables can be used to describe asthma processes or conceptions that are immeasurable but are correlated with existing variables. An example of this would be a latent variable describing breathlessness being correlated to the measureable variables forced expiratory

volume in 1 second, FEV1 and forced volume capacity FVC. Each outcome is correlated with breathless but is obviously not breathless itself but by applying latent variable models the proportion of the variance that is correlated with FVC and FEV1 can be used to describe breathless leaving other patterns of variation such as time of measurement and age of patient out of the latent variable.

Latent variables can be created for many outcome variables. The number of latent variables (factors) and the variables that correlate to them are dictated by the underlying variation patterns of the patients. Usually the latent variables are given a standard normal distribution, having mean 0 and standard deviation 1, this standardisation is carried out for ease of computation and use. Using the standardised latent variable it is assumed that the latent variable consists of one homogeneous group. For the methodology presented here the latent variable models are created using semi-parametric distributions (see chapter 5) that address the uncertainty of the latent variable distributions. This means that instead of being normally distributed the latent variable can have a much more flexible shape consisting of an infinite number of normal distributions with different means and variances, this semi-parametric technique can be used to determine sub-grouping.

Clustering methods can be split up into two types hard and soft (fuzzy). Hard clustering is when data is separated into a discrete number of categories or clusters. Each subject, or patient, is allocated to one cluster. Fuzzy clustering also describes data being classified in terms of a number of categories but this time each subject or patient can belong to one or more categories and cluster membership is expressed as a proportion between 0 and 1, in statistically models the proportion is usually obtained

as the subject's probability of being in a cluster. The semi-parametric latent variable model used in the thesis is also a fuzzy clustering method which can also be used to form a strict partition of the data to obtain a hard cluster membership (see chapters 6 and 7). The cluster membership partition can then be used to infer clinically relevant cluster outcomes for samples of severe asthma patients.

Once cluster membership is found hypothesis testing can be used to determine significant differences between the groups allowing the groups to be annotated with clinical meaning (see chapters 8 and 9). Although significant differences can be found between clusters this still does not imply distinct clusters as the clusters could make up a larger non-normal distribution. To combat this, the thesis introduces a new Bayesian statistic that is a measure of our faith in the number and distinctness of the clusters returned. This Bayesian statistic is derived from a frequentist hypothesis test for multimodality of a static distribution called the dip statistic (see chapters 6 and 7).

To demonstrate their wider applicability the semi-parametric latent variable models created for the severe asthma datasets are also used to determine sub-groups in a clinical trial for a new cancer drug (see chapter 9).

1.3 Background to asthma and asthma phenotypes

Asthma is a respiratory inflammatory disease. It is diagnosed by finding wheeze, coughing, shortness of breath, and chest tightness although these are not specific to asthma(Taylor, Bateman et al. 2008). The airway restriction seen is caused by constriction of air way smooth muscle constricting and tightening the airways,(Halayko, Tran et al. 2006) this is related to a hyper-response of the airways which can be triggered by many factors the most common being as a response to

allergens in allergic reaction pathways also associated with inflammation. The airways can be relaxed using reliever type inhaled medications, β2-adrenoceptor agonists (Halayko, Tran et al. 2006). These reverse the restriction of the air passage ways allowing responding patients to breath normally again. The inflammation in asthma is usually associated with an increase in eosinophil cells in non-severe asthma and can be treated by using preventative medications such as Inhaled corticosteroids. Both the inflammation and hyper-responsiveness aspects of asthma are much harder to quantify.

For most patients with mild or moderate asthma adequate treatment can be achieved through combination of reliever, β2-adrenoceptor agonists and controller, corticosteroid medication and good control. For severe asthma however this is not always the case. Severe asthma accounts for 5-10% of the asthma population (Holgate and Polosa 2006). Patients suffering from severe asthma have more exacerbations needing hospital admittance and tend to have poorer responses to drugs. Levels of severity can be established by spirometry measurements such as forced expiration volume in the first second, FEV1 and forced volume capacity, FVC that measures the volume of the lungs along with symptoms such as the GINA guidelines for severity classification (Salas Hernandez, Fernandez Vega et al. 2009). Most lung and thoracic societies however define severity of asthma by the level of treatment needed to obtain control, with severe asthma being defined by the American Thoracic Society (ATS) as patients who are treated with continuous or near continuous oral corticosteroids with high dose inhaled corticosteroids after being observed for a period of 6 months(Holgate and Polosa 2006).

Severe asthma is now thought of as no longer belonging to a single disease type that describes a worsening of mild/moderate asthma but rather representing an umbrella syndrome containing sub-groups with possible different pathologies (Wenzel 2003). Some of these sub-groups or sub-phenotypes are well established and are seen in clinical practice such as allergic/non allergic asthma, early onset/late onset of symptoms (Wenzel 2003). Other sub-groups however are not well established such as an obese phenotype, refractory asthma or asthma associated with high neutrophilic inflammation (Abraham, Anto et al. 2003). In these situations there is equipoise on the existence of such sub-phenotypes and whether they represent true sub-groups or represent larger non- normal distributions,(Abraham, Anto et al. 2003; Wenzel 2003).

Little statistical work has been carried out in the area of severe asthma variance analysis. Most papers centre on factor analysis, which groups together correlated variables into smaller dimensional specific latent factors that can be used to describe specific processes. Other papers concentrate on cluster analysis which groups together similar patients over a number of variables and can be used to describe different sub-phenotypes or sub-types of patients.

In its simplest form factor analysis can be carried out on the data to determine specific patterns or factors of variance within the variables. These factors can be seen as theoretical constructs or latent variables that quantify a process or aspect of severe asthma. These factors correlate with variables that are associated with the process the factor is describing (Skrondal 2004). The factors can also be described as clustering similar variables together to minimise the number of variables needed for analysis. Factors or latent variables found in severe asthma correspond to the processes body

mass, spirometry, atopy, inflammation, and symptoms, (Haldar 2008) (Moore, Meyers et al. 2009).

Clustering is used to determine specific sub groups which show small within variation but large between variations. The work that has been carried out in severe asthma and other respiratory diseases used simple algorithmic clustering techniques based on Euclidean distance measures between patients such as k-means and hierarchical clustering(Haldar 2008) (Moore, Meyers et al. 2009). These methods rely on the standardised distance between patient variables to obtain similarities between patients and thus clusters.

These methods produce partition of the data but the methods are not based on actual probability statements which could lead to subjectivity in determining the correct number of clusters (Bush and Fleming). Another issue is which variables to choose to enter the cluster analysis. Some authors use all the variables in the cluster analysis leading to the most popular pattern of variance in the data to be over emphasized (Garcia-Aymerich, Gomez et al. 2010). I.e. the one described by the most variables, usually spirometry. The better methods identify independent patterns in the data using factor/latent variable analysis and use one variable to represent each factor.

The splitting up of analysis into separate factor and cluster analysis can be considered slightly ad hoc as the data first has to be analysed through factor analysis to determine the number of factors and to determine which variables are to be used in the cluster analysis. In order to overcome the splitting of analysis's and to determine a best fitting number of clusters and a better understanding of the variability of the disease as a

whole a joint variation modelling method was created that both carries out factor analysis and cluster analysis over each factor at the same time.

These models will be used to investigate the underlying distribution of datasets of severe asthma outcomes and will be used to quantify the variation seen in severe asthma patients. These models are useful to determine which asthma variables are correlated with other asthma variables and to fully understand the underlying representation of the asthma variables. For this reason a semi-parametric distribution was used to describe each of the latent asthma variables. This special distribution allowed the latent variable to be modelled in a way free from forced parametric shapes. Usually for identification purposes a latent variable is described as being standard normally distributed (Skrondal 2004) i.e. all the subjects or patients derive from a single group. In allowing the latent variable to be semi-parametric we allow the latent variable to be described as an infinite mixture of normal distributions (Dunson 2009). This allows the latent variable to obtain any shape made from an infinite mixture of normals allowing greater flexibility. This less strict distribution means we can visualise the independent severe asthma factors better and derive a strict cluster membership or partition from information obtained from the infinite mixture model. Thus creating an advanced clustering model that is based on probabilities, obtaining the most probable number of clusters and most probable cluster membership over each severe asthma factor. In addition to this a Bayesian statistic was derived from a simple frequentist hypothesis test, the dip test, (Hartigan and Hartigan 1985) to obtain a measure of faith in the number of clusters derived i.e. a single statistic to determine whether or not the clusters predicted for each severe asthma factor are genuine clusters or represent a non-normal distribution.

Factors and clusters over factors were determined for two severe asthma datasets and are given medical/biological annotations that are in appliance with, whilst extending the severe asthma literature. The statistical methodology created was then adapted for use in a new cancer drug clinical trial to determine whether the methodology could be adapted for use in personalised medicine in a clinical trial setting, this time using a slightly different method of semi parametric modelling allowing for the different nature of the time to event variables.

1.4 Overview of thesis chapters

1.41 Chapter 2: Severe asthma variation

Asthma as a chronic inflammatory respiratory disease is introduced and defined. Also the notion of asthma severity and that severe asthma is considered not only as a severe version of mild/moderate asthma but also as consisting of several groups or sub-groups of severe asthma. The many phenotypes and asthma biomarkers are introduced and described using specific aspects of asthma as sub-headings. After this severe asthma is described using all of these aspects as a multidimensional disease and the various methods and conclusions, described from a clinical perspective, to gain inferences on these multidimensional phenotypes analyses. Predictions are made from the literature to which sub-groups could be present in the new analysis of the severe asthma datasets presented later in Chapter 8.

1.42 Chapter 3: Severe asthma datasets

The original dataset to be used was cleansed and summary statistics were taken. The original dataset however was found to be lacking in a large amount of variables coupled with a large amount of missing data in the remaining ones. There was also a large amount of data on patients that was inconsistent with patient records leading to

some patients not having a diagnosis of severity or sometimes even asthma. For these reasons the original dataset intended for analysis was not used instead two cross sectional severe asthma datasets were used for analysis. These were the Pranab Haldar dataset which was originally used for clustering using k-means algorithms in (Haldar 2008) and the Brompton blood dataset originally designed to act as a clinical phenotype dataset that could be matched with genotype data. Summary statistics and the amount of missing data are presented in this chapter along with a brief description of each variable to illustrate the properties of the severe asthma datasets.

1.43 Chapter 4: Latent variables, Factors and Cluster analysis

Clustering algorithms and mixture models are introduced as two ways of partitioning data in order to determine sub-groups in heterogeneous severe asthma data. The clustering algorithms and mixture models are reviewed in order to apply them to multivariate severe asthma data to analyse the variation found in patients. Mixture modelling was deemed the better method as this allows clusters to overlap (White, Johnson et al. 2010), while also relying on statistical modelling to gain inference on the best fitting number of mixtures. Latent variable methodology is introduced later as a way of modelling aspects of severe asthma that are not measurable. However correlated multiple outcomes that are measurable can be used in order to model severe asthma variation in variables by grouping similar variables into specific factors or latent variables. The review concludes by stating that both the latent variable/factor methodology and mixture modelling need to be combined to analyse the severe asthma variation for both patients and variables. The best method of combining these two methodologies is by modelling each factor of latent variable semi-parametrically as a Dirichlet Process Normal Mixture (Ferguson 1973; Antoniak

1974)} which looks for mixtures over the variation seen in severe asthma factors or latent variables. This allows for the latent aspects to be represented in the severe asthma datasets as groups of correlated variables which will have a more flexible distribution on them allowing for sub-phenotypes within latent variables.

1.44 Chapter 5: Semi-parametric modelling

This chapter describes semi-parametric modelling and more specifically the Dirichlet Process (DP) (Ferguson 1973) and Dirichlet Process Normal Mixture (DPNM) (Antoniak 1974) in more detail as a method of semi-parametric Bayesian modelling that allows for heterogeneity within variables both manifest and latent. Included is the statistical notation for DP and DPNM and their origins and uses in statistical fields as ways of improving fit by relaxing parametric assumptions and for use in sub-grouping data to obtain a more flexible partitioning of the data.

There are many ways to implement a DP and DPNM the methods are reviewed here towards determining the best possible statistical algorithm in order to apply the models to the severe asthma latent variables. The algorithm chosen was the Escobar and West algorithm (Escobar and West 1995) for conjugate data. Coding for the algorithm was carried out in R language (R Development Core Team 2009) an open source statistical computing language. Coding for the algorithm combined with the latent variable model for four latent variables can be found in the appendix and is explained in the second half of the chapter along with the various issues raised when coding the Dirichlet process mixture model and the statistical coding solutions used to overcome these issues. The coding of the algorithm was the main purpose of the PhD and thus took the most time to complete. First R language was needed to be understood and efficiently programmed using basic examples, then the

mathematical/statistical nature of the algorithm was needed to be understood, then a clear visualisation was needed on how the algorithm could be encoded in R. This was carried out and then improved upon. Issues while programming included long run times, multiple solutions of parameters, slow convergence, estimating convergence of parameters and determining whether mixtures were distinct or overlapping. These issues were addressed and appropriate solutions found.

1.45 Chapter 6: Simulation study using one latent variable and differing mixture scenarios

10 different latent variable heterogeneity patterns were simulated along with their continuous manifest variables in order to determine if the Dirichlet process normal mixture latent variable model DPNMLVM would be able to detect the heterogeneity and return the correct number of clusters with the correct cluster membership that explains the heterogeneity. This is carried out for 200 subjects and 500 subjects in order to test consistency within the statistical model. Also an investigation is carried out to determine if binary variables could be added alongside the normally distributed in order to determine the underlying variation, which showed that the binary variables were insufficient to predict complex continuous clustering structures.

1.46 Chapter 7: Simulation study using multiple latent variables and differing mixtures

5 different scenarios were simulated to demonstrate different patterns of heterogeneity seen for various numbers of latent variables/factors. This was carried out in order to test whether the Dirichlet process normal mixture latent variable model DPNMLVM could be used over a number of variables that displayed differing correlations with each other and thus could be explained by a number of different

independent factors again conclusions were made in order to apply the severe asthma datasets to the Dirichlet process normal mixture latent variable model.

1.47 Chapter 8: Analysis of Brompton blood dataset

The DPNMLVM is carried out using the Brompton blood dataset. The model is applied and mixtures are found and annotated for four factors. The number of clusters and memberships however were not shown to be consistent over different priors for the alpha parameter of the DPNMLVM in a specitivity analysis; the alpha prior determines the number of clusters. This was due to the Brompton blood dataset being a small dataset that allowed the priors of alpha to easily affect it. It was not analysed further due to this constraint.

1.48 Chapter 9: Analysis of Haldar severe asthma dataset

Here the DPNMLVM is carried out using the previously clustered dataset in (Haldar 2008). The model is applied and mixtures are found and annotated for four factors with some factors producing distinct non overlapping clusters. The number of clusters and memberships are shown to be consistent for three different priors for the alpha parameter of the DPNMLVM in a specitivity analysis. The truncated Dirichlet Process Normal Mixture Latent Variable model (trDPNMLVM) (Ishwaran and Zarepour 2000) was also used to determine if the two methodologies produced similar outcomes.

1.49 Chapter 10: Analysis of clinical trial dataset for a new cancer drug

The methodology is used in a clinical trial setting for a new cancer drug, using correlated survival and responder outcomes. The link functions for these outcomes are not conjugate which means the conjugate Dirichlet process normal mixture cannot be used in this case. To overcome this, the truncated Dirichlet Process Normal Mixture Latent Variable model (trDPNMLVM) (Ishwaran and Zarepour 2000) was used in

WinBUGS(Lunn 2000) to achieve a solution of the underlying distribution of these factors. Three sub-groups were found for the new cancer drug that relate to existing biomarkers not used in the analysis.

1.410 Chapter 11: Conclusion and further directions

Conclusions are made about the Dirichlet process normal mixture latent variable model DPNMLVM and the nature of heterogeneity of severe asthma. Discussion is also made about how best to carry on the research to answer both statistical and clinical important questions raised in the research that need further work to be answered. These include the application of semi-parametric structural equation modelling in order to determine regression pathways between the severe asthma factors and variables. The application of more advanced fitting criteria to determine the number of factors in the analysis, to further explore the application of binary data to the continuous data model and to allow the clustering algorithm to be adapted across all the factors rather than individual factors.

1.5 Introduction closing statement

This chapter was an introduction to the thesis as a whole including project aims, a general introduction to the concept of analysing variation in multivariate datasets by statistical models specifically in applying these models to severe asthma cohorts. The general thesis structure is outlined and a brief introduction for each chapter is described for reference and to form a logically order in which to explore severe asthma variation. We now look to the first step in applying any complicated statistical analysis to a dataset which is understanding to the best of our ability the biological/clinical question being asked of us, that of exploring and defining severe asthma variation hopefully into a number of subsets or pathologies. In order to do this a review of

severe asthma variation was carried out in order to define what variables would be important to model and to better understand the properties and clinical aspects of asthma and severe asthma. This was written up as a severe asthma review contained in chapter 2, severe asthma variation

Chapter 2. Severe Asthma Phenotypes

2.1 Chapter summary

Here severe asthma phenotypes are reviewed to determine what are the most clinically relevant when modelling the severe asthma variation. The review starts by describing asthma and then severe asthma. After the asthma introduction each aspect or phenotype is described followed by a review of existing severe asthma clustering.

2.2 Introduction to Asthma

Asthma is a chronic inflammatory disease of the lungs. It is usually diagnosed in primary care from the clinical history and presentation of wheeze, cough, shortness of breath, and chest tightness although these are not specific to asthma (Taylor, Bateman et al. 2008). Asthma affects around 300 million people worldwide and it is estimated that around 105 per 1000 persons have been diagnosed with asthma in America with the worldwide market for asthma medication exceeding $5.5 billion per year (Maddox and Schwartz 2002). Asthma is a heterogeneous disease which is thought to involve several susceptibility genes as well as many environmental features interacting with these genes. A family history is often present and supports a genetic component of asthma and that asthma may be inheritable or that related patients with asthma may have similar types of asthma or asthma phenotype.

Asthma incidence has increased dramatically in western countries over the last 10 years (Burke 2003) with asthma increasing two fold during the last two decades in Europe and now affects up to 15% of the adult population. Many hypotheses have been brought forward for this including the hygiene hypothesis, which suggests that minimizing exposure to infectious agents, by way of better personnel hygiene, coupled

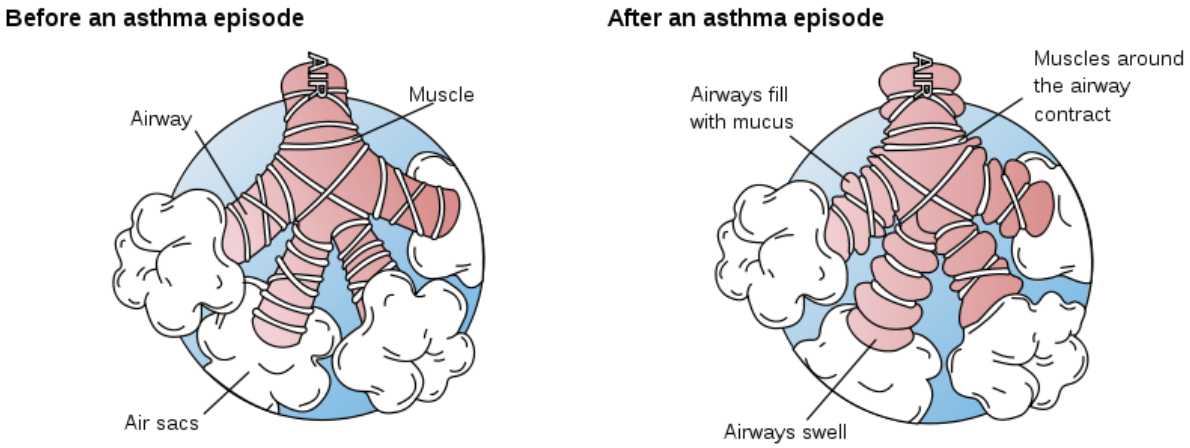
with smaller families, increases the likelihood of an atopic phenotype. However this increase prevalence could also be influenced by poorer air quality and increased air pollution affecting the airways.

Asthma is usually diagnosed based on clinical features, and is supported by the demonstration of variable airflow obstruction, see figure 2.1. The airway obstruction seen in asthma is caused by hyper-responsiness of the airways which can be triggered by many factors, the most common being as a response to inhaled allergens causing allergic inflammation. Asthma is relatively easy to diagnose in primary care and is treated initially using reliever type inhaled medications, β2-adrenoceptor agonists. These reverse the obstruction of the air passages allowing patients to breath normally again.

Other factors can affect the expression of the disease in the airways these are called triggers and can include diverse factors such as cold weather and exercise. Also gender, age, obesity and type of inflammation have a part to play also (Bel 2004). It is unclear whether these triggers/aggravators alter the expression of asthma or whether they are describing subtypes (Bel 2004)with possible different genotypes.

Such triggers, listed above, cause hyper- responsive walls to tighten causing airway obstruction and possible subsequent asthma attack. Asthma is also a chronic inflammatory disease (Faffe 2008). It was found that asthmatics had a large amount of eosinophil cells in their sputum these cells are associated with inflammation and disease severity. This discovery led to improved treatment for asthmatics by using inhaled anti-inflammatory known as corticosteroids, as eosinophils are associated with a positive response to inhaled corticosteroids.

Figure 2.1, demonstrating the effect of an asthma attack on the airways, taken from URL http://en.wikipedia.org/wiki/File:Asthma_before-after-en.sv



2.3 Introduction to Severe Asthma

For most patients with mild or moderate asthma adequate control can be achieved through combination of reliever, β2-adrenoceptor agonists and controller, corticosteroid medication. For severe asthma however this is not always the case. Severe asthma accounts for 5-10% of the asthma population (Holgate and Polosa 2006). Patients suffering from severe asthma have more exacerbations needing hospital admission, tend to have poorer responses to drugs (Wenzel 2006) and can have air-trapping and persistent airflow obstruction (Wenzel 2006). Also some patients with severe asthma tend to have structural changes in the lungs that are no longer reversible in contrast to mild/moderate asthma, as previously discussed. Patients in America with severe or difficult to treat asthma account for 50% of the total health-care costs associated with asthma (Adcock, Caramori et al. 2008). In Europe it is estimated that patients with severe asthma have an annual cost 4.8 times that of mild asthma.

One definition of severe asthma by the American Thoracic Society (ATS) workshop consensus for definition of severe/refractory asthma which is growing in popularity is

dependent on one of the following major characteristics as described in (Holgate and Polosa 2006).

- Treatment with continuous or near continuous oral corticosteroids
- Need for treatment with high-dose inhaled corticosteroids

Another paper suggests that these criteria should only be identified after being observed over a period of at least 6 months (Chanez and Wenzel 2008), due to the tendency for asthma to exacerbate. This definition along with other minor characteristics has now been adopted as the standard ATS diagnosis for refractory asthma. But defining patients severity is often complex as it can be influenced by a patient's phenotype or underlying disease activity and adherence with prescribed treatment (Taylor, Bateman et al. 2008).

Despite the use of high dose corticosteroids patients with severe asthma had the same level of eosinophils as normal asthma in one study, although neutrophils were significantly higher in severe asthma. Severity is often mistaken for difficult to manage asthma. One way to determine differences between control and severity is that asthma control can be described as the adequacy of treatment whilst severity concerns the underlying disease process and level of treatment needed(Juniper, O'Byrne et al. 1999).

Severe asthma is no longer believed to be only a general worsening of symptoms relating to mild or moderate asthma but rather that the diagnosis contains a number of sub-types. A phenotype is defined as the visible characteristics, that hopefully can be measured, of an organism resulting from the interaction between its genetic

makeup and the environment (Wenzel 2006). Sub-phenotypes of severe asthma in this case can be thought of as a sub-grouping of asthma that lies between that of different genotypes with environmental factors and different phenotypes. This sub-phenotyping or sub-grouping of asthma has been discussed in recent papers with one definition describing asthma as a cluster of related disorders (Burke 2003) with others describing the asthma definition as too broad (Wenzel 2003) or as a mixture of syndromes. Some of these sub-types of severe asthma may be similar to moderate asthma but a more severe version, while others could be specific to severe asthma.

Severe asthma may have many different triggers, types and processes these are all considered on many different levels by different authors, which is reflected in the many different types of journals containing papers on asthma and severe asthma phenotypes. The different clinical levels of the disease can be summarised using the following categories;

Physiology; Physiological Structure differences in the airways

Atopy; Asthma as a part of an allergic/atopic syndrome only

Environmental; Non allergic aggravators/triggers of asthma

Pathology/Immunology; Inflammatory cells in the lungs and inflammatory processes

Genetic Epidemiology, Genes and single nucleotide polymorphisms (SNPs) associated with asthmatic phenotypes

These categories could potentially have a large amount of overlap and sub-types seen in one group could be associated with other phenotypes in different groups, but at

present little statistical work has been carried out to obtain sub-types of asthma with only a few papers carrying out k-means and hierarchical clustering. The rest of the literature review will focus on the different sub-phenotypes found in the literature that are also measured in the datasets used later.

2.4 Immunology, Asthma as an allergic/Atopic disease only

There is a wide range of described asthma phenotypes. Although many of them may overlap most papers only consider one phenotype at a time. This is especially the case in the most common phenotype, atopic status. Atopic asthma can be thought of as asthma resulting from an allergic response to an inhaled allergen usually resulting in eosinophilic inflammation and airway obstruction.

2.4.1 Atopic Asthma

Atopic asthma is one of the best documented asthma phenotypes as it is the largest overall phenotype and is found at all levels of severity (Wenzel 2006). Atopic asthma implies that the airways are hyper-responsive to known or unknown allergens. The most common triggers of atopic asthma are dust mite, grass pollen and animal dander, although a wide range of triggers have been reported. The inflammation and airway obstruction associated with atopic airway hyper-responsiveness is due to the immunologic response to the allergen. Atopic patients allergic pathways are different then a patient with no atopy. The immune system over reacts to small amounts of allergens in atopic subjects. This was found to be the principally due to a type of T-lymphocyte known as Th2 lymphocytes. These cells produce types of proteins known as chemokines and cytokines, such as interleukins IL-13 IL-9, IL-5 and Il-4. These increase the levels of IgE antibodies, mast cells and eosinophils. They in turn control the inflammatory response associated with the atopic phenotype (Faffe 2008). This

can result in airway remodelling and constriction of the airways and if left untreated may result in an exacerbation.

Patients are usually tested for atopy by examining specific IgE antibodies in the blood or by skin prick test using several allergens. The atopic phenotype is one of the easiest to identify as samples from the lungs are not needed. It is also the most prevalent phenotype among all severities of asthma, so a large proportion of asthma research has been carried out in this phenotype including genotyping research (Cookson 2000).

At least 30-40% of severe asthmatics appear to have pathological changes inconsistent with classically described asthmatic (atopic/eosinophilic) pathology. This suggests other possibly different pathologies for specific severe asthma phenotypes (Abraham, Anto et al. 2003), (Ronmark 2007). This trend could appear because eosinophilic asthma is well treated with existing drugs, β2-adrenoceptor agonists and corticosteroids, and so patients with atopy might tend not to increase in severity as they are well treated. In contrast patients with little or no atopy are associated with persistent inflammation which may lead to structural changes involving the airways and perhaps active parts of the airways. These structural changes appear to exist well into the lung periphery out of reach of many inhaled medications (Wenzel 2003) adding to the difficulty of treatment. Patients who are atopic in the severe asthma category however tend to be very atopic for a large number of allergens and have very severe and quickly activated exacerbations (Holgate and Polosa 2006). This phenotype is more prevalent in men than women and is associated with eosinophilic inflammation.

2.4.2 Age of Onset of Disease

The age of onset of asthma disease is the age, measured in years, at which asthma symptoms first appear. Asthma onset is usually partitioned into early onset (before puberty or before 12 years) and late onset (adult onset asthma). Early onset asthma is most common in mild and moderate asthma (Ronmark 2007) and is usually associated with atopy, a history of eczema and a family history of asthma (Wenzel 2006). Late onset or adult onset asthma is more common in severe asthma (Ronmark 2007).

Late onset asthma has also been associated with a higher BMI and is more common in females then males. Adult-onset, intrinsic asthma and non-atopic asthma are now being used interchangeably to describe a sub-type defined by late onset, female predominance, higher severity, evidence of nasal polyps, (Bel 2004) and a lack of family history of asthma.

2.5 Epidemiology, non-allergic aggravators/triggers of asthma

2.5.1 Other Environmental Triggers

Allergic responses are not the only triggers of asthma attacks. There are certain triggers that can worsen asthma and can cause attacks. These can include damp or cold air, exercise (Wenzel 2006), smoking and infection with viruses and bacteria (Abraham, Anto et al. 2003). It is unclear whether a susceptibility to these triggers in itself represents a sub-phenotype of asthma or whether these conditions worsen existing sub-phenotypes that are already present (Wenzel 2006). Smoking increases asthma severity and is related to lower treatment efficacy (Van Hove, Moerloose et al. 2008) and cigarette smoke can enhance acute allergic inflammation and slows down the return to baseline.

2.5.2 Gender and Asthma

In childhood, males are more likely to develop asthma then females (Ernst, Ghezzo et al. 2002). This changes during puberty where boys are more likely to enter remission then girls and girls more likely to have severe asthma in adulthood (Ernst, Ghezzo et al. 2002). In one paper, phenotype was checked for males and females before and after puberty and the phenotype was found to be similar for both genders (Wenzel 2006) suggesting that hormonal differences at puberty did not play a major role in changing asthma phenotype.

Women seem to develop severe asthma more than men (Abraham, Anto et al. 2003) and this group is usually associated with reduced atopy and neutrophilic inflammation. It is believed that this is possibly due to sex hormonal differences as some woman seem to develop worsening symptoms at certain times of their menstrual cycle (Holgate and Polosa 2006), (Wenzel 2006) and the premenstrual period has been labelled as a trigger in one study(Abraham, Anto et al. 2003). This coupled with further research suggesting that female asthma symptoms get better during pregnancy lends extra support to a role for female sex hormones in the pathology of asthma.

2.5.3 Obesity and Asthma

Research has shown that BMI increases with severity of asthma (Abraham, Anto et al. 2003), (Ronmark 2007) and obesity is associated with an increased prevalence of asthma especially in women (Lessard, Turcotte et al. 2008). One study found that asthma control was worse in obese subjects compared to non-obese subjects and total lung capacity, expiratory reserve volume, residual capacity were lower in obese patients compared to non-obese. Obesity is also being investigated in its effects of promoting systemic inflammation, which could promote pro inflammatory hormones

that in turn could increase asthma susceptibility and severity. But at present this is still a hypothesis.

As yet it is unclear whether there is an obesity sub-phenotype in asthma or severe asthma, but this could explain the reduced response to treatments or the fact that obese patients have a higher severity of asthma. It is difficult to establish whether obese patients may have a different pathology of asthma or whether obesity acts mechanically as an aggravator as obese patients generally have lower FEV1 and lower airway closing volume and breathing near their lower airway closing volume increases airway responsiveness in asthmatic patients (Lessard, Turcotte et al. 2008). What we can say is that reducing weight in obese patients could be an important strategy for improving their asthma (Ronmark 2007) and their general health, but it is unclear whether the lung changes associated with losing weight are asthma specific or not.

2.5.4 Psychology and behavioural factors

A large amount of the control and prevention of asthma attacks is the responsibility of the patient themselves. This is due to the nature of asthma as being triggered by environmental factors. Asthma comes in attacks and the patient has to be aware of these and their seriousness and act accordingly either consistently using their preventative medication or seeking further assistance. This is where the patient's mind set can affect their asthma expression. Patients who are depressed, stressed or suffer from panic or anxiety tend to show less control over their asthma, suffer worse exacerbations and have more frequent exacerbations of asthma (Wenzel 2006), (Holgate and Polosa 2006). Thus patient self-control has to be taken into consideration when evaluating phenotypes and it is best to check that none of the patients show bad compliance with medications before assessing the phenotypes to

prevent bias in phenotyping but there are limited amounts of variables that assess compliance and control (Gamble, Stevenson et al. 2011).

It is clear that the psychological well-being of asthma patients may affect their asthmatic condition, but at present no clear psychological profile is associated with asthma (Chanez and Wenzel 2008).

2.6 Immunology/Pathology; Inflammation pathology of the cells in the lungs

2.6.1 Eosinophils

Eosinophils are inflammatory cells that are found in sputum and blood samples of many asthmatics. A high eosinophil count is a frequently seen phenotype of asthma and this reflects an eosinophilic inflammatory dominant phenotype. This high eosinophil count is usually treated with inhaled corticosteroids to treat inflammation and bring the number of eosinophil cells in the lungs down. A very high eosinophil count may be seen in poorly managed asthma, a patient having an asthma exacerbation requiring hospitalisation or a patient with severe asthma.

A possible issue when phenotyping on the grounds of eosinophil numbers is the single time point studied as it is not clear whether this time point is related to an exacerbation or is representative of the underlying inflammatory profile when clinically stable (Wenzel 2006). However for the datasets used in this thesis we only have a single time point. Thus we have to make the assumption in our analysis that the eosinophilic measurement at the baseline is similar to others measurements on that patient across time. Eosinophil inflammation is also associated with aspirin sensitivity and severe late-onset asthma. A high eosinophil cell count is correlated with basement membrane thickening (Murugan, Prys-Picard et al. 2009), a lower FEV1,

higher active symptoms and a greater likelihood for exacerbations and near-fatal events.

It is thought that one half to two thirds of severe asthma patients in the USA have persistent large airway tissue eosinophils despite adequate treatment(Murugan, Prys-Picard et al. 2009). Persistent eosinophilia appears to be more prevalent in late onset than in early onset and is a common phenotype found in patients with severe asthma.

2.6.2 Neutrophils

Neutrophils are a different kind of inflammatory cell than eosinophils. High cell counts of neutrophils in sputum may; indicate more severe airway damage, reflect inhaled steroid treatment. For these reasons the neutrophil phenotype in asthma is not clearly understood (Kaza, Bandi et al. 2007) and could even indicate a different disease such as bronchiolitis obliterans (Wenzel 2003). Neutrophils are also reported in low numbers in early onset asthma but this maybe residual inflammation associated with eosinophilic inflammation and may not indicate a distinct phenotype (Wenzel 2006). Further confusion over the phenotype can be attributed to the method of measurement i.e. sputum cell counts, as inflammatory process and cells may act differently in the distal lung with a greater number of neutrophils (Wenzel 2003) which are harder to get a sample from without invasive methods.

Neutrophils are commonly associated with other lung diseases such as chronic obstructive pulmonary disease (COPD) (Abraham, Anto et al. 2003) or Hypersensitivity Pneumonitis (Bogaert, Tournoy et al. 2009). It is thought that the airways in the severe asthmatic with neutrophil inflammation get so damaged that the asthma behaves similar to a chronic wound (Holgate and Polosa 2006) with a pathology similar to other

lung diseases as mentioned above (Bogaert, Tournoy et al. 2009). Neutrophilic inflammation is also seen in death where asthma is the cause, severe exacerbations and asthmatics who smoke (Wenzel 2006), the later possible suggesting a COPD/severe asthma cross over.

Asthmatic Inflammation is first treated with corticosteroids as these are the gold standard of treatment but in this neutrophilic sub-phenotype of patients this may not help and could even be worsening the condition as corticosteroids delay apoptosis of neutrophils (Holgate and Polosa 2006). The evidence of a distinct high neutrophil/low eosinophil phenotype is associated with structural changes found in the lung. One paper found that eosinophilic asthma is associated with thickening of the sub epithelial. Whereas in non-eosinophilic asthma it is not (Berry, Morgan et al. 2007).

2.6.3 Paucigranulocytic inflammation

This type of asthma has characterised symptoms of inflammation but no typical inflammation cells are present. It can be thought of as a similar case to that of neutrophilic inflammation suggesting that the inflammation may be out of measurements reach without invasive methods localized in the distal lung or be a different airway disease altogether as this phenotype exhibits no sub epithelial basement membrane thickening like classic asthma does. Possibly the airways may have been structurally altered to result in clinical symptoms but no inflammation as severe asthmatics usually have an increase in the amount of smooth muscle. In this case it is almost like the damage has been done, the inflammation has gone, leaving remodelled airways, which is not usually seen in classic allergic asthma. In this group up to 10% of asthma patients demonstrate poor response to glucocorticord therapy,

probably due to the lack of presence of eosinophils and have frequent exacerbations and continual symptoms (Kiley, Smith et al. 2007).

2.6.4 Refractory Asthma

In the severe asthma group there seems to be evidence of a sub-phenotype that shows little or no eosinophil cells in sputum but still retains severe asthma symptoms (Holgate and Polosa 2006) or have eosinophil inflammation that does not show any response to treatment. These patients are usually prescribed higher corticosteroids due to their symptom severity but because of the absence or small number of eosinophils in their lungs or because they show no treatment response these anti-inflammatory drugs have little or no impact, (Holgate and Polosa 2006), (Berry, Morgan et al. 2007) this sub-phenotype is known as refractory asthma. At first it was thought that this sub-phenotype was due to a defect in the patients response to corticosteroids (Wenzel 2006) or that patients had a diminished sensitivity to glucocorticosteroids in general (Abraham, Anto et al. 2003). It is now thought however that there could be another underlying process in the airway wall other than that associated with eosinophil inflammation. Cell counts have discovered large numbers of neutrophils in the cells of some of these patients (Abraham, Anto et al. 2003) (Faffe 2008) which could possibly explain the symptoms. Neutrophils are often associated with severe or irreversible lung injury, lung tissue damage(Holgate and Polosa 2006) and airway remodelling although the extent of difference is not well defined as this past remodelling may be complete and thus is not associated with inflammation anymore. The remodelling may offer some resistance to corticosteroids opposed to the physiology of normal lungs (Bai and Knight 2005).

2.7 Multivariate sub-types of airway disease

A large amount of asthma research is localised to one aspect or one phenotype of asthma, but asthma has many measurable outcomes and several different aspects as discussed above.

Few articles include all or most aspects/ phenotypes of asthma with researchers and papers concentrating on one or two aspects or outcomes of the disease. This is possibly due to a lack of multidisciplinary groups or the unavailability or reluctance to use multi variable/complex statistical analysis. There seems to be a gap in knowledge in determining sub-types across many phenotypes of severe asthma. Lots of measurable phenotypes of asthma may have common pathologies, and the amount of phenotype variation could be reduced to a few sub-types of severe asthma, but it should be observed that separating patients into strict classification via sub-type may not be useful as patients may drift between phenotypes and many phenotypes have similar characteristics, thus for some patients it may be impossible to classify them to a type of asthma, (Chanez and Wenzel 2008). In terms of an affected population however conclusions about the number of different phenotypes can and should be made.

The Early onset and late onset phenotype is well defined and was further studied by comparing outcomes of the two categories (Miranda, Busacker et al. 2004). The work in this paper backed up the theories of the late onset/early onset differences. Early onset asthma was found to have more allergic symptoms and positive skin prick tests then late onset asthma and late onset asthma again was found to have lower lung function, although eosinophilia was found in both groups and was associated with

more general asthmatic symptoms and lower lung function. Late onset also tended to have high numbers of eosinophils.

(Wenzel 2006) also suggests that asthma can be described by using the two distinct phenotypes early onset and late onset, as they both seem to differ pathologically, immunologically and epidemiologically. She later goes on to suggest that severe asthma can also be described in terms of inflammation in terms of type and presence. However 50% of severe asthma patients have very little identifiable inflammation and it has been suggested that this non inflammatory phenotype could be a type of asthma where inflammation does not have a central role.

This severe asthma phenotype was further elaborated on by Sally Wenzel suggesting that there are at least 4 sub-types of asthma with overlap of these conditions often found in patients. These phenotypes are characterised thus;

- Early onset asthma which contains higher levels of lymphocytes then late onset, a homogenous group with clear genetic and environmental allergic triggers, increase in Th2 lymphocytes and mast cells, eosinophilic disease is present and might represent a classic Th2 inflammation that is poorly responsive to steroids therapy
- Early onset asthma without the presence of eosinophils may represent a group where the inflammation had responded to steroids but the underlying disease has not or the patient's lungs have undergone permanent changes.
- Late onset asthma with worse lung function then early onset asthma coupled with higher number of eosinophils. This is a heterogeneous group with evidence for both allergic and none allergic disease. Eosinophilic asthma

includes both allergic asthma and perhaps variants of hyper-eosinophilic syndromes

- Late onset without eosinophilia share very few characteristics with the other sub-types. This may represent one or more poorly understood new types of asthma or cross over with different diseases, possibly COPD.

Although these give clear descriptions to the phenotypes there is no mention of proportions or prevalence of the phenotypes seen or quantitative measures of inflammation or atopy. However in (Wardlaw, Silverman et al. 2005) a hierarchical cluster analysis was carried out on 49 patients who were diagnosed with varying airway diseases using 8 measurable variables of patients. These were blood total IgE levels, FEV1/FVC ratio percentage, FEV1 percentage predicted, age in years, bronchodilator reversibility as a percentage of pre bronchodilator FEV1, sputum eosinophil percentage with gender and smoking status variable. In this study 4 phenotypes were found; two large groups one consisting of patients with a mainly asthma diagnosis and one with a mainly COPD diagnosis. The two smaller groups found were a COPD/asthma overlap group and an asthma group with minimal eosinophilic inflammation and low IgE levels, suggesting that cluster analysis could determine sub-groups of airway diseases. As the sub-groups found all had clinical interpretations.

This leads into a more recent cluster analysis study (Haldar 2008) concentrating on clustering 187 patients with severe asthma. In this study 6 variables were used gender, age of onset, atopic status, body mass index, sputum eosinophil count and modified JACS score, being the JACS score without the FEV1 score. This study found

four sub-phenotypes of severe asthma; these were found by utilising k-means clustering and hierarchical clustering of the asthma related variables. These were given the following clinical descriptions;

- Early onset and atopic asthma
- Obese non-eosinophilic asthma
- Inflammation predominant asthma
- Symptom predominant asthma

Clustering using hierarchical and k-means clustering is a good way to determine clusters of multivariate data, the patients can be identified as belonging to a specific type and values of proportion can be calculated. Although this is a good method the number of clusters has to be selected arbitrary basing this decision on distance between patients or cluster centres. The algorithm does not allow for patients to belong to multiple sub-types or to base the number of clusters on statistical inference or to let the sub-types overlap, see chapter 4.

Another analysis of sub-groups was carried out by (Spycher, Silverman et al. 2008) this involved a statistical analysis using latent class variables in childhood wheezing. The measurable variables from the study ranged from questionnaire data to clinical outcomes. These included categorical answers from asthma questionnaires about wheezing and clinical measures such as skin prick tests, bronchial responsiveness tests and demographics such as sex and age. Different numbers of latent classes were tried and a number of model selection criteria were used for assessing the best fitting model and thus the correct number of latent classes. Unfortunately the three types of model selection criteria used did not all concur with each other. The criteria used suggested

either three or five sub-types suggesting even this kind of statistical analysis could be too constraining for the actual phenotypes of patients. In the paper five sub-groups were suggested and described as the paper was exploratory and wanted to achieve as many clusters as possible. These were given the clinical descriptions,

- persistent cough
- transient cough
- atopic persistent wheeze
- non atopic persistent wheeze
- transient viral wheeze

It is unclear whether or not these groupings reflect the true underlying types and mechanisms of childhood wheeze phenotypes and it is also unclear whether this phenotype can be expanded to severe asthma which is mainly an adult population. For more information on these multivariate techniques please see chapter four on latent variable techniques/clustering techniques.

More recently Quantitative CT scan measurements have been used to determine if there were differences over the clinical sub-phenotypes of Severe Asthma (Gupta, Siddiqui et al. 2010). These clusters were described as

- Concordant asthma control score and eosinophilic inflammation with greater bronchodilator response
- A mainly female group with a high BMI, little eosinophilic inflammation but good asthma control
- A group with good asthma control

- And a high eosinophilic group

The clusters were not significant over the quantitative CT measurements. However when stratified for severity the CT measurements were significant.

Although sub-types of asthma have been seen in a clinical setting statistical work has been carried out to a smaller degree and factors affecting severe asthma are still not well documented (Ronmark 2007). This is possibly due to the multi-dimensional nature of the sub-types and the practicalities from obtaining relevant clinical information from the lungs and lower airways. However since the first papers on clustering technique for severe asthma have been published interest has increased in both algorithm and statistical clustering this has led to (Moore, Meyers et al. 2009) using hierarchical clustering on their severe asthma cohort. The analysis of 34 core variables produced 5 sub-groups, these were described as

- Early onset atopic asthma with normal lung function
- Early onset asthma and preserved lung function with increased medication
- Late onset asthma, non-atopic, obese women and moderate lung function
- With clusters 4 and 5 having severe airflow obstruction but they differ in their age of onset and atopic status.

This further confirms the existence of sub-phenotypes but still relies on the algorithmic clustering seen before. However it is still difficult to visualise these clusters and it is also unclear whether the clusters describe true separations or describe non-normal distributions approximated by clusters.

2.8 Conclusions

Severe asthma and asthma in general is becoming more prevalent in developed countries. This implies there is a greater need to assess and manage severe asthma phenotypes as the burden of severe asthma is likely to increase (Holgate and Polosa 2006). Although many phenotypes have been described, determining multi-dimensional phenotypes of severe asthma into specific sub-types or sub-phenotypes is still in its infancy. Many phenotypes, for example atopic and late onset asthma, are well recognised and accepted but most are still under consideration as most severe asthma research tends to concentrate on sub-types that are already established. In order to fully understand the underlying processes of all types of severe asthma more research needs to be carried out to verify new and upcoming sub-phenotypes of asthma like the low eosinophil or high neutrophil phenotype and the obese phenotype.

A statistical model for determining asthma phenotypes needs to take in to account that asthma sub-types may overlap, that patients may drift between sub-types, and which variables the sub-types are dependent on. The model should also select the most probable number of clusters and not base the decision on possible subjective clinical thinking.

Some of the multivariate sub-types that could be determined in our analysis could be;

- Inflammation predominant, early onset, atopic with High Eosinophils
- Early onset, atopic with low eosinophils
- Obese mainly female, late onset, non-atopic, low eosinophils,
- Other Severe Asthma, Late Onset, ,Low Eosinophils, High Neutrophil

Further research into these categorical sub-types could provide the information needed

- To target existing drugs to a particular phenotype
- To identify patients at high risk of exacerbation or death
- To provide clues for new drugs to be targeted to an asthma sub-type. This has already happened in the highly atopic asthma sub-phenotype and the drug Xolair.
- To better quantify severe asthma outcomes in order to carry out genome wide scans to determine genes associated with asthma and not just atopic syndrome.
- And to possibly control environmental risk factors associated with severe asthma (Burke 2003)

2.9 Chapter closing statement

We have reviewed possible phenotypes and sub-phenotypes found in the asthma/severe asthma literature and found that efforts have been made to quantify severe asthma into specific sub-groups which are perceived to be separate and have distinct pathologies and inflammation patterns. The algorithms used to determine the clusters rely on cluster numbers to be determined a prior. These methods are good at determining a strict classification of types but this might not be the case for severe asthma and patients may have similar characteristics to those in different clusters. We need a way to establish a better method of determining the correct or most probable number of clusters and to allow possible overlap of the clusters as is seen in clinic.

Clustering methods and separation of variation in datasets in general is reviewed in chapter 4. We now look at the description of the datasets to be used in chapter 3.

Chapter 3. Description of Datasets

3.1 Chapter Summary

Here we describe the two datasets used in our analysis using descriptions of the variables and summary statistics to obtain inference of the nature of the data in terms of quantity and spread of the data variables.

3.2 Introduction

Originally the dataset to be analysed came from the severe asthma clinic at the Glenfield hospital. On cleaning up of the dataset it was found that the data had a large amount of missing values. The missing data was so great that the dataset was not analysed as the data that was available showed was limited and had possible errors and mismatched with other datasets. We now look at the two datasets we have to analyse these we have given the annotation of the Haldar dataset and the Brompton blood dataset. The Haldar dataset comprises of a number of variables measured from 187 patients attending the difficult asthma clinic at the Glenfield hospital. All the patients in the Haldar dataset have a definition of difficult asthma in compliance with the ATS description. The Brompton blood dataset contains 157 patients who attended the difficult asthma clinic at the Glenfield hospital and has measurements of a number of severe asthma variables that were taken in order to link with genetic markers to test the measured variables against these as part of a bigger study.

3.3 Haldar dataset

Here we summarise the data for the Haldar dataset. A brief description of each variable is given along with its units and the percentage of missing data present in each variable, table 3.3.1. For continuous data the summary statistics mean, standard

deviation and the minimum and maximum were taken and are displayed in table 3.3.2. For binary data the proportion outcome was reported in table 3.3.3

Table 3.3.1. Description and units for the Haldar dataset with percentage of missing data for each variable.

Variable	Units	Description	% Missing
Age	Years	The age of a patient when the variables were recorded	0.00
BMI	kg/m^2	Body mass index	0.00
Onset of asthma symptoms	Years old	How old the patient was when asthma symptoms started	0.00
Eosinophils	log %	The percentage of eosinophil cells in sputum sample	0.00
Juniper Asthma Control	Score	A score representing a patients symptoms and level of control	0.00
Nitric oxide	log ppb	Nitric oxide levels exhaled, a biomarker of inflammation	0.00
Post-bronchodilator FEV1 % predicted	%	The forced exhalation volume in 1 second after taking a bronchodilator as a percentage of a patients predicted FEV1 value	0.00

BDP	mg	Dose of inhaled corticosteroid standardised by using Beclomethasone Dipropionate equivalent	0.00
Neutrophils	lg %	The percentage of neutrophil cells in sputum sample	0.00
Nigmegen	Score	The score on an a Nigmegen questionnaire	0.00
Anxiety	Score	The score on an anxiety questionnaire	0.00
Depression	Score	The score on a depression questionnaire	0.00
Pc20		A test of bronchial reactiveness	83.96
Hospital Admissions	Count	Number of admissions in previous year	0.53
Steroid courses	Count	Number of rescue oral steroid courses in previous year	6.42
ITU admissions	Count	Number of ITU admissions in previous year	0.53
Pack years	Count	Number of packs of cigarettes smoked in a year	0.00
Skin prick cat	mm above control	Skin prick test for cat allergen mm above control prick test	0.00
Skin prick dog	mm above control	Skin prick test for dog allergen mm above control prick test	0.00

Skin prick house dust mite	mm above control	Skin prick test for house dust mite allergen mm above control prick test	0.00
Skin prick grass	mm above control	Skin prick test for grass allergen mm above control prick test	0.00
Sex	(% female)	Gender indicator	0.00
Atopy (% positive)	(% positive)	Allergy indicator	0.00
Long acting bronchodilator use (LABA)	(% yes)	Variable indicating whether LABA were used or not	0.53

Table 3.3.2. Summary statistics for variables in the Haldar dataset including mean, standard deviation, minimum and maximum. (Variables logged if not normally distributed)

Variable	Mean	Standard Deviation	Min	Max
Age	43.43	15.92	14.00	84.00
BMI	28.52	6.51	18.53	64.83
Onset of asthma symptoms	20.27	18.40	2.00	73.00
Log eosinophils	0.46	1.00	-2.92	1.95
JACS	2.02	1.16	0.00	5.00
Log nitric oxide	1.63	0.36	0.790	2.36
Post FEV1 % predicted	82.06	21.06	17.00	140.00
BDP	1018.00	539.36	100.00	2000.00
Log neutrophil	1.67	0.32	-0.301	1.99
Nigmegen	17.50	13.09	0.00	56.00
Anxiety	7.3	4.66	0.00	21.00
Depression	4.66	3.84	0.00	17.00

Pc20	2.70	4.15	0.01	17.00
Hospital Admissions	1.48	1.88	0.00	10.00
Steroid courses	3.97	2.22	1.00	11.00
ITU admissions	0.36	0.74	0.00	5.00
Pack years	0.65	2.15	0.00	8.00
Skin prick cat	3.01	3.06	0.00	12.00
Skin prick dog	3.13	3.29	0.00	20.00
Skin prick house dust mite	3.79	3.50	0.00	15.00
Skin prick grass	3.43	3.27	0.00	15.00

Table 3.3.3. Summary statistics for discrete variable proportions

Variable	Proportion as percentage
Sex (%= female)	65.78
Atopy (%= Yes)	73.80
LABA use (%=Yes)	6.95

3.4 Brompton blood dataset

The Brompton blood dataset was summarised and the information obtained displayed in tables 3.4.1, 3.4.2, 3.4.3. Table 3.4.1 describes the variables and units used in the dataset with the percentage of missing data in each variable. Table 3.4.2 describes summary statistics taken from the variables including mean, standard deviation and minimum and maximum values for the continuous variables. Table 3.4.3 describes the percentage proportions of the binary variables in the dataset.

Table 3.4.1. Descriptions of each variable in the Brompton blood dataset including units and percentage missing data for each variable.

Variable	Units	Description	% Missing
Age	Years	The age of a patient when the variables were recorded	0.00
BMI	kg/m^2	Body mass index	8.28
Eosinophils	lg %	The percentage of eosinophil cells in sputum sample	10.19
Neutrophils	lg %	The percentage of neutrophil cells in sputum sample	10.83
Total IgE blood count	kU/l	The amount IgE antibodies in a blood sample	3.18
Juniper Asthma Control	Score	A score representing a patients symptoms and control	6.37
Pre Bronchodilator FEV1	Ls^{-1}	The forced exhalation volume in 1 second before taking a bronchodilator	0.00
Pre Bronchodilator FVC	Ls^{-1}	The forced volume capacity before taking a bronchodilator	3.18
Pre Bronchodilator	%	The forced exhalation volume in 1 second as a percentage of a patient's FVC value before	3.18

FEV1/FVC ratio		taking a bronchodilator	
Pre Bronchodilator FEV1 % predicted	%	The forced exhalation volume in 1 second before taking a bronchodilator as a percentage of a patients predicted FEV1 value	1.27
Post Bronchodilator FEV1	Ls^{-1}	The forced exhalation volume in 1 second after taking a bronchodilator	3.18
Post Bronchodilator FVC	Ls^{-1}	The forced volume capacity in 1 second after taking a bronchodilator	5.73
Post Bronchodilator FEV1/FVC Ratio	%	The forced exhalation volume in 1 second as a percentage of a patient's FVC value after taking a bronchodilator	5.73
Post Bronchodilator FEV1 % predicted	%	The forced exhalation volume in 1 second after taking a bronchodilator as a percentage of a patients predicted FEV1 value	5.10
BDP	Mg	Dose of inhaled corticosteroid standardised to beclomethasone dipropionate Equivalent	1.27

Steroid courses in previous year	Count	Number of rescue oral steroid courses	13.38
Oral steroid dose		Dose of oral steroid	0.00
Pack years	Count	Number of packs of cigarettes smoked in a year	1.92
Sex	% female)	Gender indicator	0.00
Atopy (% positive)	(% positive)	Allergy indicator	8.28
Long acting bronchodilator use (LABA)	(% yes)	Variable indicating whether on LABA or not	0.53

Table 3.4.2. Summary statistics for continuous variables in the Brompton blood dataset, statistics include mean, standard deviation, minimum and maximum. If variable is not normal the log of the variable is used

Variable	Mean	Standard deviation	Minimum	Maximum
Age	51.98	13.40	16.00	84.00
BMI	30.69	7.25	17.26	53.85
Log eosinophil	0.54	0.65	-0.64	1.94
Log neutrophil	1.71	0.30	0.35	2.00
Total IgE blood count	288.45	465.72	1.00	3086.00
Juniper Asthma Control	2.36	1.43	0.00	7.50
Pre FEV1	2.15	0.80	0.55	4.80
Pre FVC	3.14	1.01	0.95	6.05
FEV1/FVC ratio	0.68	0.13	0.33	0.98
Pre FEV1 % predicted	74.57	22.11	21.00	134.00
Post FEV1	2.29	0.80	0.60	4.70
Post FVC	3.28	1.01	1.25	6.15
Post FEV1/FVC	0.70	0.12	0.35	0.95

Post-bronchodilator FEV1 % predicted	78.29	22.76	1.50	138.00
BDP	78.29	22.76	0.00	4000.00
Steroid courses in previous year	3.62	3.50	0.00	16.00
Oral steroid dose	5.16	7.24	0.00	35.00
Pack years	5.41	12.22	0.00	125.00

Table 3.4.3. Proportions for discrete variables in the Brompton blood dataset.

Variable	Proportion as percentage
Sex (%= female)	63.06
Atopy (%= Yes)	66.67
LABA (%=Yes)	7.01

3.5 Discussion

Summary statistics have been taken from both datasets and the variables contained have been described. It is clear that there is a proportion of the Brompton blood dataset that have missing data. Once the variables have been selected for analysis the missing data must be removed from the specified variables to be used in the model in order to apply them. This may reduce the number of available patients used in the model thus reducing power but this is common in all large datasets.

3.6 Chapter closing statement

We have examined and described the datasets to be used in the analysis using summary statistics and semantic descriptions. We now look at the various algorithms, to use in order to quantify the variation, chapter 4, in order to model the aspects and possible grouping seen in the severe asthma phenotype, chapter 2.

Chapter 4. Latent Variable and Cluster Analysis Modelling

4.1 Chapter Summary

This chapter reviews statistical techniques used to uncover substructure in datasets by analysing the variability within those datasets. The review starts by discussing similarity measures that can be used to determine how similar a numeric object is to another numeric object i.e. Euclidean distance or correlation. These similarity measures are then described using the formal models and algorithms associated with them such as factor analysis, principle component analysis, cluster analysis and mixture modelling. The review then makes suggestions for the best way to analysis the multivariate severe asthma datasets. It is concluded that a factor analysis with infinite mixtures over each factor to determine the most probable number of mixtures in each specific aspect or factor of asthma is the best way to account for the heterogeneity found in the severe asthma datasets.

4.2 Introduction

4.2.1 Similarity measures

In order to group numerical data together we need to first establish a similarity measure that can be used to quantify similarity between individuals or variables, i.e. look for similarity between rows or columns. See figure 4.2.1

Figure 4.2.1 example of a data matrix where *Y* is a dataset containing N individuals and M variables.

$$Y = \begin{matrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{matrix} \qquad \textit{Equation 1}$$

The simplest similarity measure is the distance between two individuals. This is called Euclidean distance and is denoted *d* for a single variable say variable 1 in our example in fig 1 Euclidean distance *d* can be calculated for two individuals for variable 1, y_{11} and y_{21} by equation 2. If individuals are far away from each other *d* will be large and the points will not be similar but if the points are close by each other the distance will be small and the points will be similar,(Everitt 2001).

$$d(y_{11}, y_{21}) = \sqrt{(y_{11} - y_{21})^2} \qquad \textit{Equation 2}$$

Where *d* is the distance between individuals' y_{11} and y_{12}.

The distance measure can be easily adapted for use with more than one variable by the formula below where M is the number of variables, see equation 3.

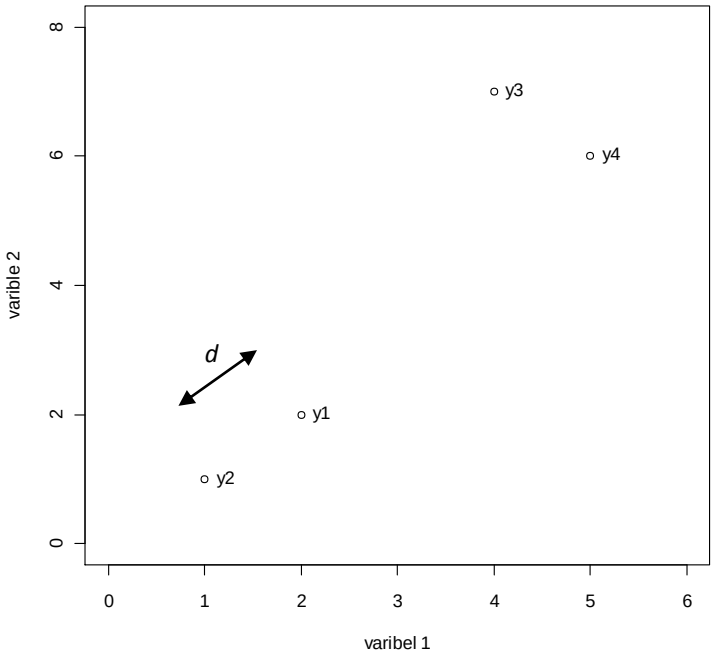
$$d(\boldsymbol{y_1}, \boldsymbol{y_2}) = \sqrt{(y_{11} - y_{21})^2 + (y_{12} - y_{22})^2 + \cdots + (x_{1M} - y_{2M})^2} \qquad \textit{Equation 3}$$

The similarity of a subset of individuals can be summarised in a matrix of Euclidean distances called a distance matrix, *D*. Thus for 4 points from fig 4.2.2 the distance matrix would be seen as *D* With the Euclidean distance, *d* between the two points $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ being shown in the entry corresponding to the first row and second column or the second row and the second row, as in equation 4.

$$D = \begin{matrix} 0.00 & 1.41 & 5.39 & 5.00 \\ 1.41 & 0.00 & 6.71 & 6.40 \\ 5.39 & 6.71 & 0.00 & 1.41 \\ 5.00 & 6.40 & 1.41 & 0.00 \end{matrix} \qquad \textit{Equation 4}$$

By viewing the similarity matrix and figure 4.2.2 it can be seen that points 1 and 2 look close together and 3 and 4 look close together. Suggesting that the simple example dataset could be explained by two clusters. Cluster 1 containing y_1 and y_2 and cluster 2 containing y_3 and y_4. The distance matrix *D* can give us indications of two lots of information from the dataset, one is an estimate to how many clusters are in the dataset and two which individuals belong in which cluster. There are many algorithms used to determine the number of clusters and the correct cluster membership of individuals, these methods will be reviewed later in the cluster analysis sub-chapter.

Figure 4.2.2, shows the Euclidean distance, *d* between the two individuals $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$



We have introduced the notation of similarity between individuals, across rows, now we look at a similarity measure between variables, across columns. A similarity measure that looks for similar patterns of variation in variables is correlation if two variables are correlated they share similar patterns of variation see figure 4.2.3. Correlation can be calculated either parametrically assuming normality, see equation 5, or non-parametrically (Comrey 1992).

Parametrically (Pearson correlation)

$$\rho_{y_1,y_2} = \frac{Cov(y_1, y_2)}{\sigma_{y_1}.\sigma_{y_2}} \qquad \textit{Equation 5}$$

Where ρy1,y2 is the correlation of y1 and y2, Cov (y1,y2) is the covariance of y1 and y2 and y1 and y2 are the variance of y1 and y2 respectively. For non-parametrically (Spearman's rank correlation) the correlation is calculated using the same algorithm but this time with the ranks of the data rather than the actual data. These similarity measures are for

continuous outcomes but can be adapted for binary variables as a McNemar test which measures the correlation for two binary variables.

Using the same methodology for the similarity measure for individuals we can create a similarity matrix for a number of correlated or uncorrelated variables see Figure 4.2.4 for an example of 4 variables which have scatter plots pairings of each and Figure 4.2.5 for their corresponding correlation matrix.

Figure 4.2.3 Graph to show correlation of two variables y_1, y_2 ρ=0.8795

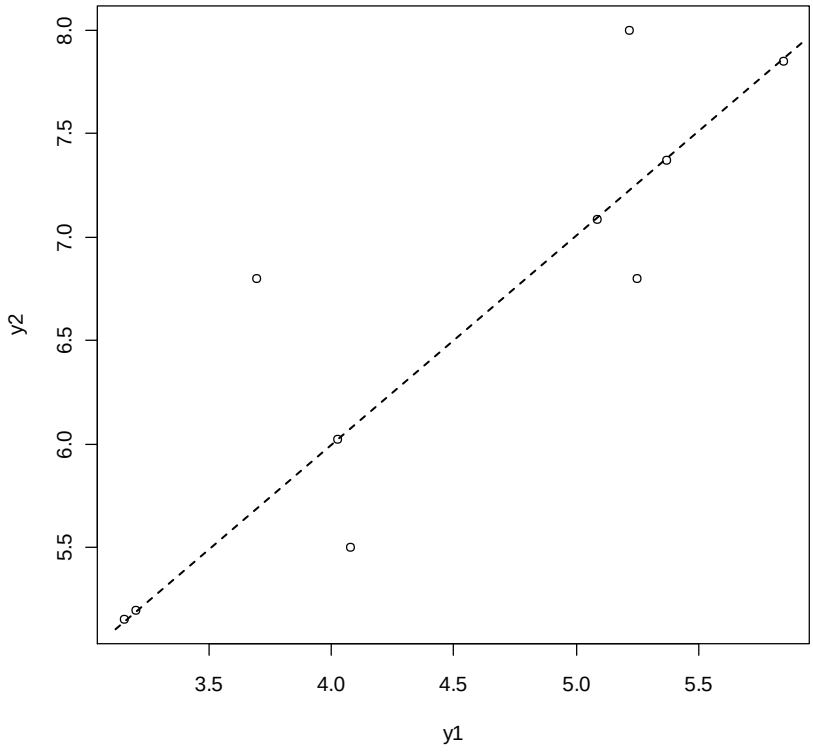


Figure 4.2.4 Matrix scatter plot for 4 variables, variable 1 and 2 are correlated and variable 3 and 4 are correlated.

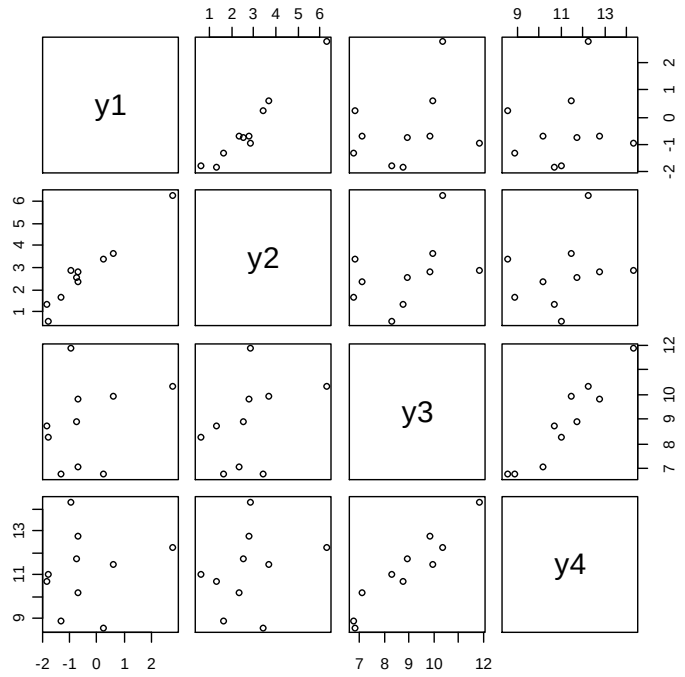


Figure 4.2.5 Example of a correlation matrix for four variables.

$$R = \begin{matrix} 1.00 & 0.97 & 0.28 & 0.13 \\ 0.97 & 1.00 & 0.41 & 0.27 \\ 0.28 & 0.41 & 1.00 & 0.95 \\ 0.13 & 0.27 & 0.95 & 1.00 \end{matrix} \qquad \textit{Equation 6}$$

We can see from the matrix scatter plot and the correlation matrix that variable 1 is correlated to variable 2 and variable 3 is correlated to variable 4, but variables 2 and 1 are not correlated to either variable 3 or 4 to a large extent. We can now see that the variables can be split into two groups we can use this information about the dataset to represent it with a lower dimensional data by only containing information for two independent variables or factors which both correlate to two manifest variables i.e. (y1,y2) and (y3, y4). The two independent correlations can be seen as a way of clustering manifest variables into latent variables or factors that represent a process that the manifest variables are correlated to, i.e. they are affected by the process. The methodology of going from a correlation matrix to

a number of factors is carried out in factor analysis and principle component analysis to obtain linear independent factors.

4.3 Latent variables

Latent variables are used to determine the underlying structure of correlated multi-variables, (Skrondal 2004). When we have a number of highly correlated variables a latent variable model can be used. Variables may be correlated because they may have the same underlying theoretical process driving them or they may be repeated measurements of the same variable using different techniques or equipment. In theory the correlated variables describe the same or similar information so we can study and quantify the theoretical process by studying the correlation of the measured variables and reducing the dimensionality. The simplest latent variable is when a measurement is taken using several different devices j each giving an approximation Y_j of the real value Z in model terms we have, as in equation 7.

$$Y_j = Z + e_j \qquad \textit{Equation 7}$$

Where Y_j are the j measured variables, e_j are the errors associated with each variable Y_j and Z is the latent variable or true value of the Y_j. In the more general model where the manifest variables are correlated with an underlying process driving them we may have a latent variable model, see equation 8

$$Y_j = \beta_{0j} + \beta_{1j} Z + e_j \qquad \textit{Equation 8}$$

Where Y_j are the j measured variables, e_j are the errors associated with each variable and Z is the underlying latent variable that is the driving force of the Y_j 's and β_{oj} ,β_{1j} are constants specific to each Y_j variable, if the Y_j variables are mean-centred then the β_{oj} become 0's.

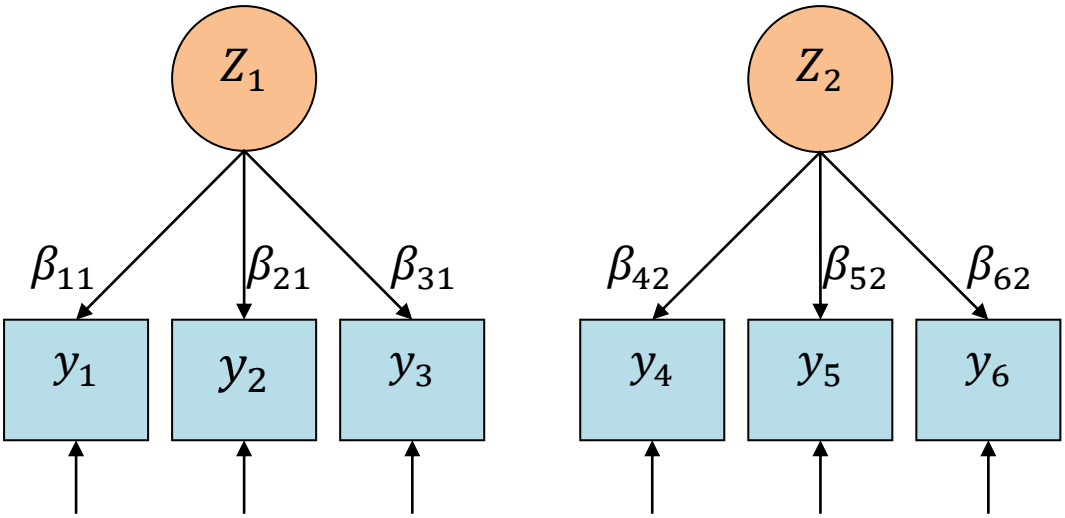
Thus the immeasurable underlying response can be quantified by using measurable variables that are correlated to it.

4.4 Factor Analysis

In Factor analysis a small number of latent variables or factors are used to describe the variation and correlation seen in a number of measured variables. Some of these variables are correlated with each other and others are not. The purpose of a factor analysis is to statistically infer the correct number of factors to use so that most of the variation in the measured variables is accounted for by these factors and to quantify the immeasurable processes so it can be used in further statistical analysis. Once the number of factors has been found we can attach meaningful definitions to them according to which variables they correlate with. In this way factor analysis can be seen as a kind of variable clustering to find certain underlying aspects of the variables. The number of these aspects is typically lower than the dimensionality of the variables, thus factor analysis also acts as a way of reducing dimensionality whilst still allowing for the variability seen in the measurable variables as the variable that correlates to the largest degree with each factor can be found and used as a representative of the factor to reduce dimensionality (Comrey 1992).

The assumptions of a factor analysis can be better explained as a figure, see Figure 4.4.1. Basically the variation seen in the measurable variables is assumed to come from underlying factors or latent variables that can't be measured directly, these factors exist and each have their own error associated with them and the specific measured variable that is correlated to it. The aim of factor analysis is to obtain these factors or latent variables so they can be used in further analysis or just to explore a dataset to obtain information on the patterns of variation.

Figure 4.4.1 Diagram of a 2 factor *(Z)* model with 6 measurable mean-centred variables *(Y_j)*. *Y* variables 1-3 are correlated with factor 1, variables 4-6 are correlated with factor 2 with factor loading $\boldsymbol{\beta_{ij}}$ for each variable i and factor j, large arrows indicate a linear relation and small arrows pointing at rectangles represent individual residuals, diagram adapted from (Skrondal 2004)



4.4.1 Factor analysis notation

We have seen how correlation can be assessed between variables and arranged in a matrix. This correlation matrix can be viewed and factors or groups of correlation can be estimated directly, this is valid when using a small amount of variables and thus factors, but with larger dataset we need a way to compute these factors and determine a variable's factor loading on each factor. The factor loading describes how well the variable correlates with the factor, factor loading have the property that the square of the factor loading is equal to the proportion of variance accounted for that factor. We need a mathematical and statistical

method to go from a correlation matrix to a matrix of factor loading. Here we present the main factor analysis equation in equation 9 (Comrey 1992).

$$y_{ki} = a_{i1}F_{1i} + a_{k2}F_{2i} + \cdots + a_{km}F_{mi} + a_{ke}E_{ik} \qquad \textit{Equation 9}$$

Where Y_{ki} is the score of the variable *Y* for person *i* on data variable *k*

F_{ji} is the factor score for factor *j* in individual i

a_{kj} is the factor loading for variable *k* on factor *j*

E_{ik} is the error score for variable *k* and individual *i*

This can be generalised to all of the individuals of a cohort dataset by using the matrix notation to form equation 10.

$$\boldsymbol{Y} = \boldsymbol{AF} \qquad \textit{Equation 10}$$

Where Y is the matrix of data variable scores which can be obtained by multiplying the factor loading A by the matrix of factor scores F. Each row of ***Y*** is the standard scores for all the data-yielding persons. Where the rows this time represent the n manifest variables and the columns represent the N patients or subjects. ***A*** is the n by m+n matrix of loading consisting of m factor loading and n error loading for the n variables and F is the m+n by N scores consisting of m factor scores and n error loading scores, equation 11.

$$\boldsymbol{Y} = \begin{matrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nN} \end{matrix} \qquad \textit{Equation 11}$$

$$\boldsymbol{A} = \begin{matrix} a_{11} & a_{12} & \dots & a_{1m} & a_{1e} & 0 & 0 & 0 \\ a_{21} & a_{22} & \dots & a_{1m} & 0 & a_{2e} & 0 & 0 \\ \dots & \dots & \ddots & \dots & 0 & 0 & \ddots & 0 \\ a_{n1} & a_{n2} & \dots & a_{nm} & 0 & 0 & 0 & a_{ne} \end{matrix} \qquad \textit{Equation 12}$$

$$F = \begin{matrix} f_{11} & f_{12} & \dots & f_{1N} \\ f_{21} & f_{22} & \dots & f_{2N} \\ \dots & \dots & \ddots & \dots \\ f_{m1} & f_{m2} & \dots & f_{mN} \\ \hline e_{11} & e_{12} & \cdots & e_{1N} \\ e_{21} & e_{22} & \cdots & e_{2N} \\ \cdots & \cdots & \ddots & \cdots \\ e_{n1} & e_{n2} & \cdots & e_{nN} \end{matrix} \qquad \textit{Equation 13}$$

To solve this equation and obtain values for the factor scores and loading we use a different version of the formula for correlation of two variables i and j that is correlation is equal to the sum of all the patient values for each variable multipled together divided by the number of patients, N.

$$\rho_{y_i,y_j} = \frac{1}{N}\sum_{k=1}^{k=N} y_{ki}\, y_{kj} \qquad \textit{Equation 14}$$

Combining the correlation equation with the factor analysis equation and assuming that each of the factors is uncorrelated with each other we obtain equation 15.

$$\rho_{y_i,y_j} = a_{y_i1}a_{y_j1} + a_{y_i2}a_{y_j2} + \cdots + a_{y_im}a_{y_jm} \qquad \textit{Equation 15}$$

In matrix form the correlation matrix R equals the matrix of factor loading multiplied by the transpose of the matrix factor loading, AA', see equation 16. Thus we have an equation from a correlation matrix to a matrix of factor loading. The error values are not usually computed as these are usually not needed as it is the factor loading that are of interest in the analysis. There are many methods to split the correlation matrix *R* into two matrices which are the

transpose of each other. One easier way is to allow *A* to be the same dimensionality as *R*, but this would defeat the object of trying to obtain factors of less dimension then the variable space, (Comrey 1992). In reality *R* is approximated to the best ability in order to obtain the lowest dimensionality of *A* as the measured variables will always have an amount of error variation with them (Skrondal 2004).

$$R = A\,A' \qquad \textit{Equation 16}$$

The methods used to find *A* usually start by adding just one factor and maximising the factor loading associated with this factor, A_1 to allow for the most variation i.e. the factor loading squared. This variation is then removed from *R* as described in equation 17.

$$R_1 = R - A_1 A_1' \qquad \textit{Equation 17}$$

A second factor is added and again the variation associated with the factor is maximised to determine the factor loading A_2 to determine the best factor loading this is again subtracted from R_1 to get R_2, as in equation 18.

$$R_2 = R_1 - A_2\,A_2' \qquad \textit{Equation 18}$$

This process is repeated until there is not enough variation to create a factor and an error matrix is created to encapsulate the rest of the variation

4.4.2 Selecting factors

The correlation matrices *R, R1, R2, R3*, all have the special property of being symmetric positive semi definite (PSD) which means the value at y_{12} is the same as the value at y_{21} i.e. that variable 1 is correlated to variable 2 independent of order and that y_{11} is equal to 1 i.e.

a variable is perfectly correlated with itself thus it can be said that if a matrix is PSD it can be seen to be made of a number of eigenvectors i.e. the factors, each eigenvector is associated with an eigenvalue which corresponds to the variance explained by the factor if this eigenvalue is greater than 1 then it is said to be practical significant and explains a large amount of variation but eigenvalues less than 1 are not practically significant and only explain a small amount of the variation seen in the data, this rule is called the Kaiser criterion and is often used to choose the best fitting number of factors. To determine the eigenvalues we use the determinant of the matrix A using equation 19.

$$\det(A - \lambda I) = 0 \qquad \textit{Equation 19}$$

Where det is the determinant of the matrix. A is the matrix of factor loading and λ are the eigenvalues of A.

We have looked at the formula needed to calculate factor analysis we now look at the different types of factor analysis including principle component analysis as a special case of factor analysis. There are two main types of factor analysis, exploratory and confirmatory. Confirmatory requires prior knowledge to create an existing framework between the latent variable model and the measurable variables. Exploratory factor analysis is used when no prior knowledge or hypothesis is used and the aim of the analysis is to find out the hidden framework of correlation and underlying processes.

4.4.3 Exploratory factor analysis

Exploratory factor analysis aims to determine the optimum number of factors to explain the variance seen in a finite number of measured normally distributed random variables. Usually the emphasis of the analysis is to quantify hidden or immeasurable variables or to find links between variables, but it can also be used for dimensionality reduction similarly to

principle component analysis, see later chapter 4.4.6. Unfortunately due to the exploratory nature of the analysis, the resulting factors of the analysis are not often backed up in other studies (Skrondal 2004) one reason for this may be that the factors are possibly being over-fitted to individual datasets.

Once found the factors are usually given meaningful annotations and realisations of the dataset can be made. The factors can prove difficult to interpret in some models however and it is worth noting that the factors are specific to the outcome variables used and the nature of the factors could change when more outcome variables are added. Exploratory factor analysis is a useful method when little is known about the variables or their variation but a better method if prior information is available is that of confirmatory factor analysis as the structure is specified before.

4.4.4 Confirmatory factor analysis

Confirmatory analysis is when the variables are assumed to relate to specific known factors and the knowledge of the variables and factors is specified in the model framework, thus it can act as a hypothesis test to assess if the confirmatory factor model is correct or not. This method is more rigid then exploratory analysis but is useful when trying to assess latent variables with known associations in previous publications.

4.4.5 Principle Component analysis

Principle component analysis (PCA) is a statistical model that tries to explain the variance seen in a number of continuous variables by selecting the smallest amount of components that can explain the maximum amount of variability. Although very similar to factor analysis and indeed principle component analysis has been used to carry out factor analysis but PCA has a different set of assumptions. In factor analysis factors or latent variables are assumed to exist but are immeasurable directly so the factors are created by regressing on variables

that are correlated with the factor or process to measure these underlying processes. In PCA however the factors or components found are just ways of reducing the dimensions of the data but still allow for the full variance in the data.

So in factor analysis the factors drive the outcome variables and the factors describe processes but in PCA the variables drive the components to create a smaller dataset where the components don't represent underlying processes or latent constructs. So they can be seen as almost equivalent methodologies that were designed in different scientific principles, factor analysis in psychometrics to quantify immeasurable psychological processes (Skrondal 2004) and principle component analysis to reduce dimensionality of data to speed up data mining algorithms (Dunteman 1989).

The factor analysis methods presented here are fine when the data is homogeneous and normally distributed. However in many cases as with the severe asthma data there is a large amount of heterogeneity and this heterogeneity needs to be accounted for within factors. Also how can binary variables be added to the factor analysis as binary factor analysis,(Ansari and Jedidi 2000) exist but methods for combining both binary and normally distributed variables are scarce, these are typically calculated using Bayesian techniques as the integrals are difficult to solve and involve assuming that the binary variables have an underlying normally distributed distribution driving them.

Methods for accounting for heterogeneity in multivariate data are now reviewed, in an attempt to adapt the latent variable/factor model into one that can cope with non-normal/ possible multi-modal data, in which we would like to infer possible evidence of sub-groups, mixtures or clusters.

4.5 Cluster analysis

Cluster analysis determines a number of sub-groups, k of patients which reduce the within cluster variation but maximise the between cluster variation. To obtain this partition of the data a similarity measure is needed to establish the similarity between patients over variables. There are many different distance metrics that can be used with Euclidian distance being the most popular. Euclidean distance is described by the formulae below for the two individual's y_1 and y_2 as in chapter 4.2.1 and equation 20

$$d = \sqrt{(y_1 - y_2)^2} \qquad \textit{Equation 20}$$

In order for the similarity measure to work effectively the variables to be used need to be standardised as to not allow scales of the variables to effect the partitions. This can be easily done by converting the variable into the standard normalised scale by applying the z-score technique (Everitt 2001).

The methods for cluster analysis rely on distance measures to determine subgroups of multiple normally distributed outcomes and can be thought of as mathematical algorithms rather than statistical models. The mathematical algorithms are commonly used by computer scientists and are relatively straight forward to carry out in contrast to statistical models which are used by statisticians to infer mixtures based on statistical inference (Mirkin 2005). There are many ways to use the distance matrix to determine a cluster partition of the data the two main ones being k-means clustering which states the number of clusters to be found a prior and achieves the best partition of the data using the k number of clusters iteratively and hierarchical clustering which either starts with every patient being in their own cluster and slowly merges them together iteratively until

eventually they all end up in the one cluster or does the reverse of this where the patients start out in one global cluster and slowly segregate till eventually they are in there own separate cluster, which is a less popular method (Mirkin 2005). Once the algorithms have converged the best-fitting number of clusters can be obtained this is done either by comparing statistics for the differing number of clusters in the k-means clustering method (Everitt 2001) or by comparing the partitions in hierarchical clustering using a diagram called a dendrogram (Mirkin 2005).

Statistical models can be used to sub-group data by using mixture models. A mixture model can be used to describe whether a single variable distribution or multivariable distribution can be better explained using a mixture of distributions. This mixture modelling is usually carried out by applying the mixture models with increasing numbers of mixtures and then using statistical model selection techniques to determine the number of mixtures the data fits best (Kuo, Aggen et al. 2008), (Lubke 2008). This method is commonly used by statisticians, but problems can arise from different model selection techniques choosing possible different solutions of differing numbers of mixtures (Lubke 2008). Further problems can be incurred when dealing with multivariate mixtures, as is seen in the severe asthma dataset, with increasing dimensionality comes extra parameters to evaluate and longer running times in software.

4.5.1 Clustering algorithms

In this section we summarise the clustering algorithms used to determine partitions of multivariate data into subsets based on a distance measure. These clustering algorithms are usually separated into two type's k-means and hierarchical.

4.5.2 Hierarchical clustering

The hierarchical algorithms work by merging or separating subjects or clusters of subjects. The algorithm starts by allocating every subject to its own individual cluster. The data is then merged a subject at a time by selecting a different subject with the smallest distance away from the other points in the cluster. The process is carried out iteratively until all the clusters and subjects are in the same group. Alternatively you can do the reverse by placing all the data points in the same cluster and separating the clusters to minimize the overall distance between clusters until every data point is in its own cluster. The steps of the former hierarchical process can be described thus,(Everitt 2001).

Start Assign every individual in its own cluster C_i for i= 1,2,…, N where N is the number of points

1. Find the nearest pair of distinct clusters say C_i and C_j using the distant measure merge these two clusters and remove C_j as the two clusters are now both represented as C_i due to the merge.
2. If there is only one cluster left stop otherwise repeat step 1 until there is only one cluster.

The resulting steps of the algorithm can be represented in a figure called a dendrogram which describes the partitioning of the data, by depicting individuals along the bottom of the dendrogram and U shapes joining vertically as and when the clusters converge. The distances of the U shapes between clusters is the distant of each cluster compared to the other cluster. This depends on the hierarchical method used. The different methods of hierarchical clustering come from the way the nearest pair of clusters C_i and C_j is calculated. The different methods each have different assumptions with them and look for different

patterns within the data, several reviews have compared the difference methods with simulated continuous data (Milligan 1980) and binary data (Hands and Everitt 1987) with directions suggested for the best method to determine the types of clusters wanted and the type of data to be used in the analysis. Although there are methods for clustering both binary and normally distributed variables together this is not commonly carried out in standard software as some transformation of one type of data is needed.

4.5.2.1 Single linkage

Single linkage is the simplest form of hierarchical clustering and it defines the distance between two clusters as the shortest distance between two points one each from each cluster (Everitt 2001). All pairings of individuals, one from each cluster are checked and the pairing that has the shortest Euclidean distance is merged. This is a simple and efficient algorithm and can reveal complex patterns that other clustering algorithms cannot find such as lines or circles of data, but this kind of clustering is rare in clinical data as usually we assume that the variables are normally distributed or made up of a mixture of populations that are each normally distributed all be it with a small amount of error. For this reason it is not often used in clinical applications as the shapes of the clusters can be undesirable.

4.5.2.2 Complete linkage

This is the opposite of single linkage clustering with clustering dependent on the distance of the pairs of individuals, one from each cluster, that are the furthest away from each other. This again has the same disadvantages as single linkage as the clusters can again become odd shapes. Both single linkage and complete linkage are better methods for looking for chains of data rather than spherical or multi-normal distributed clusters.

4.5.2.3 Group-average linkage

This hierarchical clustering method uses all the distances created between all of the pairings of two individuals one from each cluster and then averaging them into a mean creating a group averaging distance between two clusters.

4.5.2.4 Centroid linkage

A centroid is described as the centre of a cluster it can be seen as the mean of the data for the cluster. In centroid linkage the distance is measured form the mean or centroid of the data in the cluster to the centroid of another cluster using the standard two step algorithm. This is a good method if the clusters that are being merged are similar sizes otherwise the bigger cluster may dominate the smaller cluster in representing the new merged centroid (Everitt 2001), this is especially seen at the beginning stages of the algorithm. In some data cases the median could be used instead of the mean to better represent the data, when merging non similar sized clusters.

4.5.2.5 Wards method

The Wards method try's every possible paring of clusters, like the other methods and the two clusters whose combination result in the smallest increase in a statistic called the information loss are combined. This information loss function is calculated using the error sum of squares criterion (EES) (Ward 1963), see equation 21. The EES is calculated for each partition of clusters and the best partition is chosen this is the merge with the smallest EES. This cluster methodology gives the clusters a spherical shape, similar to a multi-variate normal distribution.

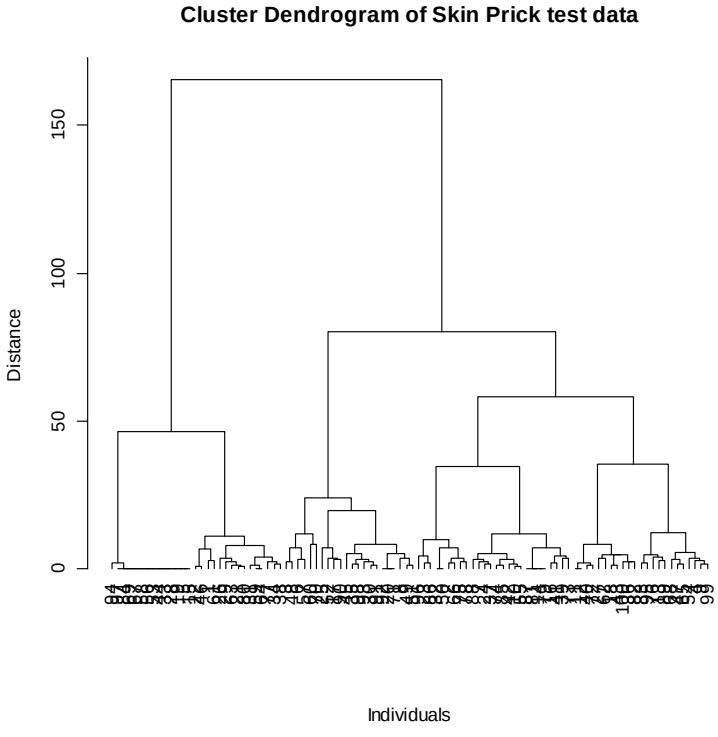
$$EES_{one\ group} = \sum_{i=1}^{n} (y_i - mean(y))^2 \qquad \textit{Equation 21}$$

$$EES_{three\ groups} = EES_{group\ 1} + EES_{group\ 2} + EES_{group\ 3} \qquad \textit{Equation 22}$$

Where y_i are the individuals in a cluster with n individuals.

The distances can be displayed in a dendrogram, which display the hierarchical nature of the data and the distance change between different clusters or in the case of wards algorithms the difference in the ESS, see Fig 4.5.2.5 for an example of a dendrogram based on four skin prick tests for cat, dog, grass and dust mites on 100 individuals.

Figure 4.5.2.5 Dendrogram to show clustering of skin prick test data, the best number of clusters can be found by comparing the differences between cluster solutions. The biggest difference is between the solution for cluster 2 and 3 indicating that the cluster 2 solution is the best fitting solution.



Hierarchical analysis is a good method to partition data and has been used extensively in practice as it can give a clear diagram (dendrogram) to explain the nature of the different number of clusters. The x axis represents the individuals in the datasets and the y axis distance or similarity of the clusters. Corners represent merging of clusters the vertical distances between cluster merges are the differences in distance between the cluster merges. The bigger the distance between the different merging clusters the better the representation of the previous cluster.

Hierarchical clustering methods however do not allow the clusters to overlap and the cluster membership does not derive from probability statements or statistical inference and due to the hierarchical nature of the algorithm once data points enter a cluster they can then not be removed to be attached in a different cluster which might be better suited, leading to possible mis-clustering (Hopke and Kaufman 1990). Another effect of the algorithm is that it arranges the data in to a hierarchy this makes sense with some data such as pedigree and protein structures, where the differences could be down to hereditary it does not always match every dataset where there is no such hierarchy. It is also difficult to determine the optimum number of clusters; this can be inferred by slicing the dendrogram at a particular level, usually the one with the biggest difference, For the skin prick test data example this largest distance would be for 2 clusters and it is worth noting that in some studies this is chosen arbitrarily to obtain an exploratory cluster analysis.

The question remains which algorithm is the best to use this is dependent on the data if looking for spherical clusters where the clusters are based on measurements with some error associated with them then wards algorithm or centroid linkage is probability the best as these methods are based on average distance measures allowing for errors. These

methods are closer to statistical models then the other linkage hierarchical methods which are more useful to detect clusters that look for chains of data or outliers. Most scientists are looking for spherical clusters which are based on measurements with errors for this reason wards method is often used in the literature and can be seen to achieve the best results in comparison with other hierarchical methods in simulation.

4.6 K-means clustering

Similar to hierarchical clustering the k-means algorithm is based on a distance measure that creates a partition of the individuals. But unlike hierarchical clustering k-means clustering requires the desired number of clusters to be specified a prior before carrying out the clustering (Mirkin 2005). K-means clustering is an iterative process that is initialised by picking k points in the multivariable space as the initial k cluster representatives or centroids, mean of the cluster. Followed by an iterative process to determine the best fitting partition of the data into the k clusters, using a distance measure. The iteration process is separated into two steps (Mirkin 2005).

Start Centroids are chosen

1. Each subject is assigned to its closest centroid, similar to centroid clustering resulting in a partition of the data,
2. The cluster centroid/mean is recalculated using the new data points in the clusters.
3. When the centroids no longer change the process has converged (Wu, Kumar et al. 2008). This stopping rule is called the Minimum distance rule.

The closest subject is decided using a distance measure and again the Euclidian distance is the obvious choice. Convergence however can sometimes reach a local optimum partition instead of the correct solution however as the algorithm is sensitive to starting values (Wu,

Kumar et al. 2008). It is important that good starting values should be used by inspection of the data although this could lead to problems with prior beliefs. Determination of the k number of clusters is usually carried out by carrying out different numbers of clusters and applying some criterion using a statistic related to mean distance or variation of the clusters for deciding which cluster number, k is used such as a cost function (Ahmad and Dey 2011) or the number of clusters can be assumed from a previous hierarchical clustering. Alternatively if carrying out an exploratory analysis the clusters are chosen in line with clinical thinking, as is the case for the previous analysis of severe asthma data (Haldar 2008) although this was loosely based on hierarchical clustering.

Both the k-means clustering and hierarchical clustering can be sensitive to outliers as they both rely on distance as a similarity measure and not probability statements. Despite these drawbacks both methods are easy to implement and comprehend.

4.7 Finite mixture models

The k-means clustering algorithm can be thought as a hardened version of a statistical mixture model. In both procedures the number of groups or clusters/mixtures are stated a prior, the difference is in mixture models the clusters are distributions, i.e. the normal distribution with specific parameters, i.e. mean and standard deviation parameters that are dependent on the individuals in the cluster (Garcia-Escudero, Gordaliza et al. 2010). Individuals have a probability of belonging to a specific distribution mixture. In k-means clustering however the individuals are only allowed to be in one cluster/mixture (Wu, Kumar et al. 2008). Individuals are put in a cluster depending on the distance from their centroid, where as in mixture models the individuals are associated to the cluster/mixture by their probability of being in the distribution associated with the cluster/distribution with specific

mean and standard deviation parameters. If the data is well partitioned, i.e. the clusters are far away from each other then the mixture modelling and the k-means algorithm obtain the same partition but if the data is more fuzzy, i.e. it is hard to say whether subjects may belong to either cluster, especially where the clusters overlap as is normally found in reality, the different methods obtain different solutions which could lead to over simplification of the data and misclassification of subjects when using the k-means clustering methods as the mixture models allow for this overlap.

The model for a uni-variate mixture model is described below and can easily be extended to multivariate cases. Emphasis of the mixture model review here has been on normal and multivariate normally distributed variables but other distributions can be used such as Weibull distributions for survival outcomes (Entink, Fox et al. 2011) or Bernoulli mixtures for binary data(Kim 2003).

The mixture model equation can be described for two mixtures as in equation 23

$$f(y) = p.f_1(y) + (1-p).f_2(y) \qquad \textit{Equation 23}$$

Where *f* is the distribution of variable *y.* f_1 is the first mixture distribution and f_2 is the second mixture distribution, *p* is the probability associated with the first mixture distribution and 1-*p* is the probability associated between the second mixture distribution each f_i distribution will have its own set of parameters.

The formula for two mixtures can be adapted for k mixtures as in equation 24 and 25.

$$f(y) = \sum_{j=1}^{k} p_j \, . f_j(y|\theta_j) \qquad \textit{Equation 24}$$

$$where \sum_{j=1}^{k} p_j = 1 \qquad \textit{Equation 25}$$

Where *y* is the data variable, $f_j(y|\theta_j)$ is a mixture distribution with parameter θ_j and p_j is the probability of the individual belonging to the mixture distribution $f_j(y|\theta_j)$. For normally distributed variables $f_j(y|\theta_j) = N_j(y|\mu_j, \sigma_j)$. The formulae representation above lays the foundation for the maximum likelihood estimation of the possible many parameters in the model. The log of the likelihood is maximised to determine the best fitting solution of the parameters in the model, equation 26.

$$\log Liklihood = \sum_{i=1}^{N} \ln[\sum_{j=1}^{k} p_j . f_j(y_i|\theta_j)] \qquad \textit{Equation 26}$$

Where *N* is the number of individuals, *k* is the number of mixtures, y_i is the value of the data variable for individual i, $f_j(y|\theta_j)$ is a mixture distribution with parameter θ_j and p_j is the probability of the individual belonging to the mixture distribution $f_j(y|\theta_j)$.

The equation can then be solved in the usual way by using the expected maximising (EM) algorithm or by Bayesian estimators of the parameters. This creates one of the problems with mixture models as the above formulae is fine for implementing small numbers of mixture components and small numbers of patients but with large amounts of both, the methods to evaluate the parameters can become too computationally expensive and other issues arise when using multivariate data as assumptions have to be made about the covariance of the mixtures and restrictions our often implemented on these so that solutions are obtainable, such as independence of variables and mixtures.

Once the mixture models have been successfully implemented, the best fitting number of mixtures can be assessed by comparing different models in the standard way using model fitting criterion such as BIC, AIC or maximum likelihood. Each different model has a different number of mixtures. Unfortunately different model criteria can determine different numbers of mixtures and it is often not clear which model criteria to use contributing to uncertainty over the number of mixtures found, which is a common problem in mixture model analysis.

4.8 Infinite mixture models

We have summarised the procedures for finite mixtures where the number of mixtures are changed with each model and then the models are compared using model selection criteria to determine the best fitting number of mixtures, but another method would be to have one statistical model infer the best fitting number of mixtures. This can be obtained by allowing the number of mixtures in a model to be infinite and to have them converge to the most probable solution. This can be formulated by extending the mixture formula above to infinite mixtures, equation 27 and 28, where the parameters are as listed previous.

$$f(y) = \sum_{j=1}^{\infty} p_j \,.\, f_j(y|\theta_j) \qquad \textit{Equation 27}$$

$$where \sum_{j=1}^{k} p_j = 1 \qquad \textit{Equation 28}$$

The infinite mixture model can be implemented only in a Bayesian framework in a Dirichlet Process Mixture Model as the EM algorithm cannot be used for imputation as the infinite mixture model uses a special Dirichlet distribution based prior called the a Dirichlet process or Dirichlet process mixture. Originally created by (Ferguson 1973) and adapted for mixtures by (Antoniak 1974). These can be difficult to implement and to converge due to the large amount of parameters needed to evaluate but have been carried out recently thanks to advances in Markov Chain Monte Carlo Techniques (MCMC) (Escobar and West 1995). MCMC techniques are a way of sampling from parameters to create a distribution for the parameters thus obtaining a mean for each of the parameters. They also allow a number of complex statistical models to be carried out and solutions obtained that cannot be found when using the EM algorithm. For full details of Bayesian statistics, MCMC algorithms and the Dirichlet process mixture models see the next chapter on semi parametric models, chapter 5. As the Dirichlet process normal mixture is a semi parametric distribution it can take any shape that can be created using an infinite amount of normally distributed mixtures thus it is a lot more flexible then other distributions that are only restricted to certain distributional shapes which can be unrealistic.

4.9 Combining latent variable with cluster analysis and mixture modelling

Usually mixture models are carried out over a number of similarly distributed variables i.e. all normally distributed or all binomial variables but if latent variables are used to reduce the dimensionality and infer on the clustering of the variables, these two can be combined with mixture modelling allowing a mixture analysis to be carried out over each latent variable or factor, creating a lower dimensional space speeding up the algorithm, while allowing for patterns of variation in variables and individuals. These techniques are reviewed here.

Factor analysis reduces the dimensionality of the variables and describes the heterogeneity of the variables allowing the factors to be annotated with biological reasoning, but the heterogeneity of the patients within the factors still has to be accounted for, this is usually assumed to be normally distributed for ease of computation. Many simple methods have been devised to address this issue one is to carry out a cluster analysis on the individuals by using the factors or a representative variable with strong correlation to each factor, i.e. the highest factor loading (Haldar 2008) . Once the independent variables/factors are obtained they can be used for cluster analysis or a statistical mixture modelling could be applied to the factors.

This clustering could be considered a slightly ad hoc method as in many factor analysis's the variables are assumed to be normally distributed and by running a cluster/mixture analysis afterwards we assume that this is not the case, but it is possible to carry out non-parametric factor analysis and thus check for mixtures in the factors. This has been carried out in practice as it is easily done in two steps first obtain factors and then second obtain clusters from the independent factors, because the factors are independent the mixture modelling

or cluster analysis become easier models to compute as the covariance of the data variables can be kept at 0.

Another method applies a more complex model which has the mixtures being distributed over each of the factors combining factor and mixture model analysis in one complete statistical model, a mixture/factor model. This is a less ad hoc process as there is no break in methodology but is not that common in the literature as the models are difficult to create and suffer from problems arising from identifiability issues, as in some situations the model has trouble determining whether the variance should be described using a factor or a mixture, thus creating several solutions (Lubke 2008). To express the model as a formula we start with the factor analysis equation as before for *k* variables, *j* factors and *i* subjects in equation 29.

$$y_{ik} = a_{k1}F_{i1} + a_{k2}F_{i2} + \cdots + a_{kj}F_{ij} + E_{ik} \qquad \textit{Equation 29}$$

Where Y_{ik} is the score of the variable *Y* for person *i* on data variable *k*

F_{ij} is the factor score for factor *j* in individual *i*

a_{kj} is the factor loading for variable *k* on factor *j*

E_{ik} is the error score for variable *k* and individual *i*

Instead of having the factor scores normally distributed with mean 0 and standard deviation of 1 we adapt the factors for mixtures, obtaining the new model with the same parameters as the other models, equation 34.

$$y_{ik} = a_{k1}F_{i1} + a_{k2}F_{i2} + \cdots + a_{kj}F_{ij} + E_{ik} \qquad \textit{Equation 30}$$

$$F_j = F_{ij} \qquad for\ all\ i\ 1\ to\ N \qquad \textit{Equation 31}$$

$$F_1 = \sum_{l=1}^{L1} p_{1l} \cdot f_{1l}(y|\theta_l) \qquad \textit{Equation 32}$$

$$F_2 = \sum_{l=1}^{L2} p_{2l} \cdot f_{2l}(y|\theta_l) \qquad \textit{Equation 33}$$

$$F_j = \sum_{l=1}^{Lj} p_{jl} \cdot f_{jl}(y|\theta_l) \qquad \textit{Equation 34}$$

Where F_{ij} is the factor score for factor *j* in individual *i*

F_j is the factor variable containing all of the factor scores F_{ij} for factor *j*

f_{jl} are the *l* mixtures over factor *j* with parameter Θ_l

p_{jl} are the probabilities associated with being in mixture *l* for factor *j*

Lj are the number of mixtures associated with factor *j*

The issues involved in implementing these models are similar to the problems with mixture models which are that different model criterions infer different models, this does not imply that one model is right and the others are wrong rather that there are many possible answers to the problem the model specifies. Selection of the best model is even harder in the factor mixture models because we have to make a joint decision on both the number of factors and the number of mixtures on each factor. This could lead to many models being specified and possibly many models having similar fitting (Lubke 2008).

4.10 Cluster analysis to diagnose sub-groups in medical databases

Cluster analysis has been used to diagnose several sub-groups using medical databases this is either done by clustering several outcomes of patients with a known diagnosis of a disease from a clinician to invest the underlying heterogeneity (Haldar 2008) or individuals from both disease cases and controls can be used to determine if a disease cluster is returned after analysis (Wardlaw, Silverman et al. 2005). Usually a factor analysis is carried out first and then a cluster analysis using the factors or the highest factor loadings from the factor analysis so that the data aspects or factors are equally represented for clustering (Haldar 2008), (Folkerts, Nagel et al. 1990).

Practically all of the methods to partition individuals are carried out using cluster analysis rather than mixture modelling as the analysis is usually carried out in SPSS and until recently mixture modelling could not be carried out in SPSS. This means that the papers have some of the failings of cluster analysis that were mentioned previously that it is difficult to determine the correct number of clusters and some of the clusters are chosen to back up previously reported hypothesis or that lots of clusters are looked at to determine an exploratory cluster analysis. Most of the diseases that have undergone cluster analysis are diseases that are characterised by symptoms which are likely to include a large amount of heterogeneity as the diseases definition are not based on specific biomarkers but symptom reporting. This can lead to possible different pathologies having the similar disease definition. Generally this is exactly the reason why cluster analysis has been carried out in severe asthma. Apart from asthma disease datasets other diseases that have undergone cluster analysis include neurological diseases such as Parkinson's disease (van Rooden, Heiser et al.) and depression in the medically ill (Guidi, Fava et al. 2011) . The purpose of each cluster analysis remains similar, to verify existing or hypothesised clusters, to

determine possibly new mechanisms and/or to determine possible new drug targets for disease types.

4.11 Conclusions for modelling severe asthma datasets

In modelling asthma we would expect the variables associated with asthma to be correlated with each other if they share similar pathology. If the disease was homogenous we might expect all of the variables to be highly correlated with each other. In reality the data are not that simple and the variables can be split up into sub-sets of variables that are highly correlated with each other, i.e. a multiple latent variable model. These different latent variables can also be seen in the literature review as different aspects of asthma variability i.e. sputum cell counts and asthma symptom scores may only be related to a specific aspect of asthma say underlying inflammation and skin prick test scores and IgE levels may be correlated to another aspect or latent variable of asthma say atopic status (Leung, Wong et al. 2005), see the severe asthma literature review in chapter 2 for more details of these aspects. This factoring of asthma implies that several latent variables should be used to describe severe asthma, each one describing a different aspect. Multiple latent variable analyses are commonly seen in social and psychometric research where typically the process that needs to be quantified is difficult to measure by other means. In this area of research the latent variables are called factors and the statistical modelling to be carried out is called factor analysis.

In the severe asthma dataset the reason we are using these reduction techniques is primarily to determine if sub-groups of subjects exist in the reduced dimensions and secondary to see what variables are correlated to the factor so that the sub-groups and the factors can be given annotations to better understand severe asthma as a disease process.

For these reasons factor analysis, latent variable modelling is what is required rather than principle component analysis.

My research analysis is to determine the reasons for the variation seen in severe asthma patients and it is unclear whether the variance in the severe asthma dataset is due to the existence of sub-groups, individual differences of severity or different aspects/factors of asthma. For these reasons and because of the lack of prior knowledge as most of the literature is aimed at childhood non-severe factor analysis (Leung, Wong et al. 2005) I chose to use exploratory factor analysis to combine the many variables that are related/correlated to severe asthma but which it is unclear whether the variables are correlated with each other or not.

Computational algorithms for multidimensional clustering such as k-means and hierarchical clustering can be limited and do not rely on statistical inference. A better method would be to use a statistical model to determine the number of groups, mixture models can be used to determine the best possible number of groups but these rely on model selection techniques. Choosing a correct mixture model however is not a standard problem and different model selection techniques can produce different number of optimum clusters for the same data. A better method is to allow the statistical model to choose the number of mixtures this is carried out using an infinite number of mixtures. To carry out such an analysis Bayesian inference is used to implement a Dirichlet Process Mixture. This is the subject of the next chapter, chapter 5.

The Dirichlet Process Mixture also allows more flexibility in evaluating the latent variable values and can also be used to infer sub-groups. Strict clustering can be obtained from its MCMC output that is similar to that of k-means clustering but without relying on distance

that limit the k-means and hierarchical algorithms. So to summarise the severe asthma dataset will be analysed using a factor analysis with an infinite number of mixtures on each factor.

4.12 Closing statement

Methods of dividing variation within data allowing for both variable variation and individual variation has been reviewed. The best method to allow for variation in variables is factor analysis and the best method to allow for variation in individuals is mixture modelling as cluster algorithms are not based on probability statements making it harder to determine the best-fitting number of mixtures/clusters. Even mixture modelling has its disadvantages one of these being difficulty in choosing a model fitting criteria that is the best suited for mixture modelling. This can be overcome by allowing an infinite mixture model that converges to the best fitting number of mixtures. These can be difficult to implement however. We will analysis the severe asthma datasets by applying a factor analysis/latent variable model that has an infinite number of mixtures over each factor to allow for variation within individuals in that specific factor or latent variable. Infinite mixtures and how to implement them are covered in the next chapter on semi-parametric models, chapter 5.

Chapter 5. Semi-Parametric Modelling

5.1 Chapter Summary

This chapter reviews non-parametric methods for Bayesian inference concentrating on methodology and implementation of Dirichlet process normal mixtures (DPNM). This is reviewed in order to use the DPNM to determine the density of latent variables. Using a semi-parametric technique such as a Dirichlet process normal mixture to calculate the density of the latent variable allows the latent variable to be described without the confines of parametric assumptions as an infinite number of mixtures thus allowing the true nature of the latent variable to be shown and if multi-modal, subgroup allocation can be achieved to find a partition of the data. Implementing Dirichlet Processes and Dirichlet Processes mixtures however is more complex than using a normal distribution to describe a latent variable.

Semi-parametric techniques are introduced and Bayesian inference described in general, reasons for choosing a semi-parametric prior are discussed and the mathematical notation and properties of the Dirichlet process (DP) and the Dirichlet process mixture (DPM) are stated. In the later part of the chapter issues concerning implementation of the Dirichlet process mixture (DPM) model are raised and solved. These issues include non conjugate priors, ease of computing, differing Monte Carlo Markov Chain (MCMC) computing strategies, and benefits of complete versions over approximations, and speed of computing. A description of how the model was coded in R using both R and C programming languages is also included.

5.2 Introduction

In parametric analysis the data are assumed to belong to a family of simple parametric distributions that are used to describe the data as a probability distribution. The use of parametric distributions is only correct however if the assumptions for using the parametric distribution shapes hold. If this is not the case the shape of the data may be over-simplified leading to limitations in the scope and type of inferences being made (da Silva 2007), possibly leading to increased errors, a bad fitting model and bad predictions for future data (Dorazio, Mukherjee et al. 2008). If the data represent a single group with individuals sharing similar characteristics with a single mode than parametric assumptions maybe fine as we assume the data is homogeneous. If the data is heterogeneous however, the data may possibly contain different types or clusters of patients, this heterogeneity maybe lost under parametric assumptions. Non-parametric and semi-parametric distributions do not rely on shapes to fit the data they show the true nature of the data allowing for possible heterogeneity but not excluding the possibility of homogenous distribution.

Non-parametric distributions as opposed to having no parameters as the name suggests actually have many parameters but none that dictate the shape of the distribution, allowing the distribution to fit the data better. Semi-parametric distributions have a mixture of parametric shapes governed by a non parametric distribution thus it would be better for heterogeneous data to assume a non-parametric or a semi-parametric distribution, allowing the data to be freer of parametric assumptions (Ghosh 2003),(Dey, Muller et al. 1998).

Non-parametric Bayesian priors can be used to obtain non-parametric posterior inference similar to that of frequentist non-parametric estimation. The most used

non-parametric distribution is the Dirichlet process. This non-parametric prior was devised by (Ferguson 1973) and creates a discrete random probability measure over the space of all probability measures. The Dirichlet process was developed further to include probability measures over an infinite amount of mixtures (Antoniak 1974) if the mixture distributions used are continuous parametric distribution functions then the model can be described as an infinite mixture of continuous parametric distributions allowing for continuous data to be freed from the confines of using just one parametric shape. This Dirichlet process mixture (DPM) is a type of semi-parametric prior as it has both non-parametric and parametric assumptions. Both the Dirichlet process prior and the Dirichlet process mixture prior have become the most popular non-parametric techniques thanks to MCMC computational algorithms allowing the Dirichlet process and Dirichlet process mixtures to be fitted. MCMC algorithms were first developed for the Dirichlet process and Dirichlet process mixtures by (Escobar and West 1995). The use of the MCMC methods freed the methodology so that they could be calculated in a wide number of varying Bayesian hierarchical models.

5.3 Bayesian techniques

Bayesian techniques allow statisticians to quantify their own beliefs into probability statements and determine how these prior statements are updated by sample data *y* to formulate posterior beliefs. This posterior inference can provide information that can be used to estimate parameters in a statistical model. To better formulate this into a statistical model we can think of the prior, data and posterior as belonging to parametric distributions, where the prior distribution is driven by our prior beliefs. As with beliefs these priors can be narrow/informative or vague/non-informative we

could believe some data should have a mean of 5 and put a probability of one on this. Alternatively we can be vague and state that we are not sure of the value and give a distribution for the parameter with equal probable values between -100 and 100. Either way it is clear that a prior *p(θ)* can be formulated about a random variable θ. We can also assume that θ can also be a parameter in a distribution. So for random variable *y* with distribution *p(y|θ)* with parameter θ probabilities can be determined about θ from data *y* formulated by *p(θ |y)* where data *y* is fixed and θ dependent on *y* is determined by equation 35

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \qquad \textit{Equation 35}$$

Where *p(θ |y)* is the posterior, *p(y| θ)* is the probability of the data as a function of *y*, *p(θ)* is the prior assumptions of θ and *p(y)* is the normalising constant needed to make *p(θ |y)* into a proper probability measure, this is a form of Bayes rule (Lee 1997).

Where the normalising constant is defined for the continuous case by equation 36

$$p(y) = \int p(\theta)p(y|\theta)dy \qquad \textit{Equation 36}$$

And for the discrete case defined by equation 37

$$p(y) = \sum_{y} p(\theta)p(y|\theta) \qquad \textit{Equation 37}$$

When *y* is constant and *p(y| θ)* is a function of the data,y, we call p(y|θ) the likelihood of θ given data *y*, see equation 38

$$p(y|\theta) = l(\theta|y) \qquad \textit{Equation 38}$$

By maximising the likelihood or log likelihood we can find the best or most likely solutions for the parameter θ. Note this value is without the influence of the prior. Using the prior however allows us to incorporate prior information and use computational techniques that are very flexible.

By omitting the normalisation constant and replacing *p(y|θ)* with the likelihood we have the formula in equation 39

$$p(\theta|y) \propto l(\theta|y)p(\theta) \qquad \textit{Equation 39}$$

Thus posterior inference can be achieved over the space of parameter θ (Lee 1997). Ideally for ease of computation we would prefer that the likelihood and the prior information when multiplied together created some function that could be manipulated into a standard statistical function/distribution that is recognisable and easy to sample from. For these reasons it is best to use conjugate priors these are priors that have the same type of distribution as their related posterior. The choice of distribution however also has to be a valid way of describing the prior information and data and not just used for convenience of computing the posterior (Lee 1997). When a prior, a function of θ, is conjugate then when it is multiplied by the likelihood also a

function of θ the posterior is of the same parametric family as the prior. This method holds well when only one parameter is to be inferred on. But when many are needed this can involve complex priors.

This was overcome by using Monte Carlo Markov Chain (MCMC) techniques that sample from the posterior of one parameter and then use this parameter as a constant in a sample from a different parameter posterior creating an iterative chain (Gelfand and Kottas 2002). This can better be demonstrated using a series of equations for an example of three parameters. Consider Bayes rule for posterior probability in equation 39 and consider the parameters represented by θ are three parameters $\theta_1, \theta_2, \theta_3$.

$$p(\theta_1, \theta_2, \theta_3|y) \propto l(\theta_1, \theta_2, \theta_3|y)p(\theta_1, \theta_2, \theta_3) \qquad \textit{Equation 40}$$

Instead of predicting new values all at once involving complex distributions we can predict the values of $\theta_1, \theta_2, \theta_3$ one at a time, by keeping the others constant. We start by inialiasing the parameters with starting values, $\theta_{10}, \theta_{20}, \theta_{30}$. We then update for parameter $\theta_1$by sampling a value θ_{11} from the posterior $p(\theta_1, |y)$ which is based on the likelihood $l(\theta_1|\theta_{20}, \theta_{30}, y)$ of $\theta_1$dependent on the data y and the starting values of the other parametersθ_{20}, θ_{30} and the prior $p(\theta_1)$ of θ_1, see equation 41.

$$p(\theta_1, |y) \propto l(\theta_1|\theta_{20}, \theta_{30}, y)p(\theta_1) \qquad \textit{Equation 41}$$

Next we sample from the posterior of $\theta_2$to obtain new value θ_{21} using the starting value of θ_3, θ_{30} and the new value of θ_1, θ_{11}, see equation 42.

$$p(\theta_2, |y) \propto l(\theta_2|\theta_{11}, \theta_{30}, y)p(\theta_2) \qquad \textit{Equation 42}$$

Similarly parameter 3 is sampled see equation 43

$$p(\theta_3, |y) \propto l(\theta_3|\theta_{11}, \theta_{21}, y)p(\theta_3)$$ *Equation 43*

This cycle concludes the first iteration and has obtained new values for the three parameters. The process is run again, Predicting θ_1 from the new values of θ_2 and θ_3 see equation 44 and likewise for parameters θ_2 and θ_3, see equation 45 and 46.

$$p(\theta_1, |y) \propto l(\theta_1|\theta_{21}, \theta_{31}, y)p(\theta_1)$$ *Equation 44*

$$p(\theta_2, |y) \propto l(\theta_2|\theta_{12}, \theta_{31}, y)p(\theta_2)$$ *Equation 45*

$$p(\theta_3, |y) \propto l(\theta_3|\theta_{12}, \theta_{22}, y)p(\theta_3)$$ *Equation 46*

The process is called Gibbs sampling (Lee 1997) The process continues iteratively and the chain is run for a long time and the samples are stored as it is run. We can use the samples form the parameters to plot the posterior distribution of the parameters as a histogram that describes our data. The chain only depends on the previous parameters values, the data and the prior for the parameters so the chain should forget what starting values it started with and these values can be cut leading to a chain that has converged to the correct solution. Once convergence has been achieved inferences can be made from the saved sampled parameters and statistics obtained from the samples for the parameters (Gelman, Carlin et al. 2004) achieving a mean and standard deviation for the parameters.

As explained before these priors are usually specified to be a parametric distribution whaving parameters θ to aid conjugatecy in order to compute easily and efficiently, but these assumptions could be invalid when using heterogeneous data and semi-parametric priors may offer an alternative that creates a better fit when no prior knowledge of the distribution is known. For the asthma data we are using latent variables to describe the variance seen in a number of variables associated with severe asthma unfortunately it is impossible to determine the parametric shape of latent variables so it is a logic step to assume a semi-parametric distribution on the latent variable (Lee 1997).

5.4 Non-parametric methods

Non-parametric probability distributions don't rely on parametric assumptions that restrict the distribution shape of the data. Thus non/semi-parametric distributions can be used to better fit the data in a more relaxed way. Classical non-parametric distributions can be difficult to incorporate into hierarchical models to avoid this I used Bayesian inference so that complex relationships in the data could be used. The problem now equates to how to select a prior to mimic non-parametric inference (Lee, Lu et al. 2008). In non-parametrics, there are no parameters but in Bayesian statistics by its nature has to have parameters to determine posterior inference on. To get round this difficulty a special prior needs to be used. Instead of the prior being a probability measure or distribution over a parameter space, the non-parametric prior will have to be a probability measure over a space of probability measures or probability distributions (Muller and Quintana 2004). The non-parametric prior will therefore be infinitely dimensional. The problem now lies in how to formulate an infinitely dimensional prior and how to calculate the posterior from the prior and

likelihood in an effective and relatively straight forward way. The Dirichlet process was introduced in (Ferguson 1973) as a non-parametric prior that is also a discrete probability measure that satisfies the two conditions for a good prior stated in (Ferguson 1973).

That the support of the prior distribution should be large

Posterior distributions given a sample of observations from the true probability distribution should be manageable analytically

Although the Dirichlet process is an effective non-parametric prior it is discrete. This led the Dirichlet process to be adapted for mixtures of distributions by (Antoniak 1974) and by allowing these infinite mixtures to be continuous parametric distributions allows the data to be described as a continuous prior, with fewer restrictions of shape, called a Dirichlet Process Mixture. These mixture priors are called semi-parametric as they allow the data to be described as an infinite non-parametric mixture of parametric distributions. With the application of MCMC methods for implementing calculation of the posterior by (Escobar and West 1995), (MacEachern and Muller 1998) and (Ishwaran and Zarepour 2000), the Dirichlet process and the Dirichlet process mixture have become the most popular non-parametric/semi-parametric priors. These priors can be used in hierarchical models for better model flexibility.

5.5 Dirichlet processes

Dirichlet processes are the most commonly used non-parametric technique in Bayesian statistics. A Dirichlet process can be thought of as a discrete probability measure over the space of probability measures. We will now give a formal statistical definition of Dirichlet processes.

The Dirichlet process is a special distribution that when sampled from creates a new probability distribution itself. The Dirichlet process has two parameters α and G_0 and can be described in equation 47

$$Dirichlet\ process = D(\alpha, G_0) \qquad \textit{Equation 47}$$

The G_0 parameter is called the location parameter and it determines the position of a point in the distribution on a space Θ (Ferguson 1973). If the space Θ is the real line *R* which is very common in distributions, then the G_0 parameter states the points on the real line that are in the Dirichlet process. The amount of probability associated with the points in G_0 is described by the precision parameter α and the data being distributed. To visualise this we can think of a continuous line of numbers with G_0 being a selection of these numbers or points and then the alpha parameter working out the probability associated with these along with the actual data.

For large datasets the alpha parameter has little effect but for smaller ones alpha can have an effect on the distribution, this is similar for most priors in Bayesian models. The Dirichlet process is a conjugate prior meaning that if data is given a Dirichlet process prior then the posterior distribution of the parameter space will also be a Dirichlet process. To go from the prior to the posterior we can see this as estimating an unknown probability distribution posterior G (da Silva 2007). Consider G_o as our prior guess at G based on a sample size of n, usually all the other data, G can either be from G_o with probability p_0 or belong to a different point contained in G_n with probability (1- p_0), see equations 48 and 49.

$$G = p_0 . G_0 + (1 - p_0) . G_n$$ *Equation 48*

$$(1 - p_0) . G_n = p_1 G_1 + p_2 G_2 + \cdots$$ *Equation 49*

$$where\ p_0 = \frac{\alpha}{(\alpha + n)}$$ *Equation 50*

$$and\ p_1 = \frac{n_1}{(\alpha + n)}$$ *Equation 51*

$$and\ n_1 + n_2 + \cdots = n$$ *Equation 52*

p_i is the probability of belonging to the point G_i and is thus calculated using how many points there are within G_i as they representative the n_i points. So G is a mixture of the prior guess G_o and the empirical distribution function of the G_n. If α is large compared to n then little weight is given to the data conversely when α is small. E.g. when α =0 G is given by G_n which is the classical nonparametric Bayes estimate.

This gives a flexible non-parametric prior but this prior is discrete leading to a discrete posterior, if continuous data is used then the Dirichlet process has to be adapted, the adaption was first carried out by (Antoniak 1974) creating a Dirichlet process mixture.

This new process has been described as a Chinese restaurant process. Imagine a Chinese restaurant with an infinite number of tables. A person from the sample, G comes into the restaurant and gets seated at a table G_1. The next person to come in has a choice either to sit with the person on their table G_1 and will choose this with probability p_1 or sit on their own table selected from G_0 with probability p_0 if the new person chooses to sit on their own then their table becomes G_2. The third person comes and either sits with one of the other two people with probability p_1 or p_2 or sits on their own with new probability p_0. The parameter alpha determines whether they sit on their own or not, where as the more people who sit round a table the more likely another person will be to sit at that table. In our case the tables are distributions or mixtures the people are patients and the Chinese restaurant is a patients underlying asthma.

5.6 Dirichlet Process Mixtures

Dirichlet processes are by their nature infinite discrete random measures, for this reason they are limiting to discrete variables, but by allowing the G and thus G_n and G_o in the Dirichlet process to be a continuous parametric distribution instead of just a point then we can obtain a semi-parametric continuous prior that can be described using equation 53

$$F(y) = \int f(y|\theta) dG(\theta) \qquad \textit{Equation 53}$$

$$G(\theta) \sim D(\alpha, G_0(\theta)) \qquad \textit{Equation 54}$$

Thus F can be described as a mixture of $f(y|\ \theta)$ distributions with a Dirichlet process prior on the random mixing measure G covering a mixture of possible parameters θ.

Equivalent models can be formulated by extending finite mixture models to infinite mixture models (da Silva 2007) see previous chapter on finite mixture models, chapter 4. The most common parametric mixture in a Dirichlet process mixture is a normal distribution we call this a Dirichlet process normal mixture (DPNM) model. To establish the mixture connection we will look at an example of a parametric Bayesian model for determining the individual means of a set of data and then the more relaxed semi-parametric model using a Dirichlet process normal mixture.

A formal hierarchical Bayesian model for data variable *Y* can be laid out in equations 55, 56, 57.

$$Y_i \sim f(\theta_i) \qquad \textit{Equation 55}$$

$$\theta_i \sim G(v) \qquad \textit{Equation 56}$$

$$v \sim H(a, b) \qquad \textit{Equation 57}$$

Where the Y_i is individual data i for the one manifest or measurable variable Y distributed parametrically via distribution f with corresponding individual parameter θ_i which has prior distribution *G* with hyper parameter *v* which has prior *H* who's parameters are fixed values *a* and *b*. For normally distributed data with known

variance θ_i becomes the mean parameter and the above formula translates to the new equations 58, 59, 60.

$$Y_i \sim N(\theta_i, \sigma^2) \quad \textit{Equation 58}$$

$$\theta_i \sim N(\mu, V) \quad \textit{Equation 59}$$

$$\mu, V \sim N(\mu_0, V_0).IGamma(a_0, b_0) \quad \textit{Equation 60}$$

Here we assume that the individual means θ_i of Y_i can be normally distributed, but this condition can be relaxed by substituting the parametric *N(μ, V)* in the normal case with the non-parametric Dirichlet process normal mixture prior $D(\alpha, G_0)$ thus obtaining for the normally distributed variable *Y*

$$Y_i \sim N(\theta_i, \sigma^2) \quad \textit{Equation 61}$$

$$\theta_i \sim G \quad \textit{Equation 62}$$

$$G \sim D(\alpha, G_o) \quad \textit{Equation 63}$$

$$G_o \sim N(\mu_i, V_i) \quad \textit{Equation 64}$$

$$\mu_i, V_i \sim N(\mu_o, V_0).IGamma(a_o, b_o) \quad \textit{Equation 65}$$

$$\alpha \sim Gamma(a, b) \quad \textit{Equation 66}$$

Thus data can be given semi parametric priors. The Dirichlet process priors allow the distribution of data to be modelled with uncertainty of parametric shape which also includes inference on clustering of the nature into infinite mixtures, while still being flexible enough to be incorporated into hierarchical models. For these reasons the Dirichlet process normal mixture is ideal for monitoring the distribution of a latent variable in order to answer both our primary and secondary research questions which are

1. is severe asthma variation represented by a continuous variation in severity or does it exist due to sub-groups or clusters of patients with similar characteristics and
2. If these sub-groups exist can they be quantified and clinical inferences made about the sub-groups.

5.7 Dirichlet Process and Dirichlet Process Mixture uses

Dirichlet process and Dirichlet process mixtures are used in Bayesian hierarchical models for three main reasons. These are to determine the underlying shape of a distribution this is when the primary interest in the model lies with determining the distributional shape of a variable (Xing, Jordan et al. 2007), which would be useful in density estimation of manifest asthma variables or in allowing latent variables relating to asthma to have a more relaxed distribution. To improve fit of a model where parametric distributions may not be valid (Dorazio, Mukherjee et al. 2008) or to determine sub-groups due to the clustering nature of the Dirichlet process, Dirichlet process mixture, (Brown 2008) another useful property which could be applied to the severe asthma datasets.

To increase model fit Dirichlet processes and Dirichlet process mixtures are used to model variables where assumptions of parametric distributions are difficult or

impossible to prove or are simply false causing bad predictions (Dorazio, Mukherjee et al. 2008) or biased results (Kleinman and Ibrahim 1998). This type of application includes random effect models such as in meta-analysis. In the meta-analysis, the random effect variable which is usually assumed to be normally distributed can be given a Dirichlet process to capture possible heterogeneity between studies, (Chung, Dey et al. 2002) 2002}, (Kleinman and Ibrahim 1998). Blocked models similar to random effect variables, where the block effect is distributed semi-parametrically have also been used (Bush and Fleming 1996).

Another case where a non-parametric distribution is useful and would possibly increase goodness of fit is the distribution of errors in a model these again are usually assumed to be normally distributed which is not always the case, this was used in an instrumental variable model (Conley, Hansen et al. 2008).

The clustering nature of a Dirichlet process and Dirichlet process mixture is a useful property and can be used to infer subgroups of subjects this is useful for carrying out mixture modelling/clustering or machine learning. Dirichlet process models are a better method then k-means clustering for determining mixtures and clustering multivariate data as the number of clusters does not have to be specified prior to applying the methods. Dirichlet Process models have been used in bioinformatic applications such as gene expression analysis, protein sequence analysis (Brown 2008) and classification of multivariate data by prior specification such as genetic abnormalities {Hionoff, 2005}. Machine learning applications include matching words with images where the words can be seen as variables and the images as clusters associated with a subset of the words (Barnard, Duygulu et al. 2003).

Examples of determining shapes of distributions from variables include determining latent variable distributions. These are used in all areas of research where latent variables are used, examples include; determining latent fraud detection (Xing, Jordan et al. 2007), studying Psychological Cognitive Behaviour (Navarro, Griffiths et al. 2006), latent home owner insurance claim behaviour (Braun, Fader et al. 2006) and studying heterogeneous populations of animals (Dorazio, Mukherjee et al. 2008) to determine possible evidence of sub-groups.

5.8 MCMC techniques for implementing a Dirichlet Process Normal mixture

We can use MCMC techniques to fit a Dirichlet process (DP)/ Dirichlet process mixture (DPM). First we split the parameters of the latent variable model above in to two groups those that are needed to implement the Dirichlet process mixture and those that are used for the other parametric parameters. The parametric parameters not associated with the Dirichlet process mixture can be easily implemented by Gibbs sampling, the Dirichlet process mixture parameters however need to be treated slightly differently because of their infinite semi-parametric nature. The MCMC techniques to implement the semi-parametric process can be described as belonging to two sets of algorithms, Marginal or conditional, (Dey, Muller et al. 1998), Marginal algorithms integrate out parameters that are not needed in the analysis to reduce the parameter space thus speeding up the algorithm by making the MCMC algorithm less complicated. The conditional algorithms use all the parameters iteratively to obtain estimates for all the parameter space. These different methods are reviewed here to determine the most applicable to a Dirichlet process mixture latent variable model.

5.8.1 Marginal MCMC algorithms

Sampling from the Dirichlet Process can be carried out in different ways. The first MCMC algorithm was devised by (Escobar and West 1995). This method used the polya urn scheme representation of the Dirichlet process mixture (Blackwell and Macqueen, 1973) to produce the joint posterior distribution of Y_i's. The G distribution has been integrated out making the algorithm marginal, although this is an example of a marginal algorithm it is still conditional as the parameters are dependent on other parameters indicated by the symbol | . θ |Y, ς, λ indicates that θ is dependent on the other parameters Y, ς, λ as in equation 67.

$$[\theta|Y,\varsigma,\lambda] \propto \prod_{i=1}^{n} f(Y_i|\theta_i,\varsigma) \ . \ \frac{\alpha\, G_o(\theta_i|\lambda) + \sum_{k<i} \delta(\theta_i|\theta_k)}{\alpha + i - 1} \qquad \textit{Equation 67}$$

Where $f(Y_i|\theta_{i,},\varsigma)$ is the probability distribution function of Y at θ_i and $\delta(d\theta_i|\theta_k)$ is the simple distribution which is a point mass on $\theta_{k.}$ G_0 is the prior distribution for the location parameter of the Dirichlet process, α is the precision parameter, λ denotes the hyper parameters on G_0 and ς represents all parametric parameters not involved in the Dirichlet process mixture. This is similar to the Chinese restaurant example of the how the Dirichlet process mixture chooses the posterior points each point has a choice of a new point from G_0 with probability α/(α+i-1) multiplied by the likelihood of the point Y_i being in the probability distribution $f(Y_i|\theta_{i,},\varsigma)$ and the probability of being in another individuals distribution i.e. $f(Y_i|\theta_{j,},\varsigma)$ multiplied by how many points are already in that distribution divided by(α+i-1) i.e. as in equation 68.

$$\delta(d\theta_i|\theta_k)/(\alpha + \mathrm{i} - 1) \qquad \textit{Equation 68}$$

Multiplied by by the likelihood of the individual Y_i being in another individual probability distribution $f(Y_i|\theta_{j,},\varsigma)$.

By extending the formula to the limit $\alpha \to \infty$ i.e. we always chose a new distribution we have equation 69.

$$[\theta|Y,\varsigma,\lambda] \propto \prod_{i=1}^{n} f(Y_i|\theta_i,\varsigma)\ .G_o(\theta_i|\lambda) \qquad \textit{Equation 69}$$

This represents the base prior with no clustering effect where all θ_i come from the initial prior G_o, but as alpha gets smaller the θ_i are largely based on the data that make up the other θ_k's that are close to the θ_i. We can then apply the Gibbs sampler on the posterior distribution to form the posterior as in equation 70.

$$[\theta_i|\{\theta_k, k \neq i\}, Y_i, \varsigma, \lambda] \sim$$
$$q_o\ .\ G_{lim}(d\theta_i|\lambda, Y_i, \varsigma) + \sum_{k<i} q_k\ .\delta(\theta_i|\theta_k) \qquad \textit{Equation 70}$$

Where G_{lim} is the base prior and

$$q_o \propto \alpha \int f(Y_i|\theta_i,\varsigma)\ .dG_o(\theta_i|\lambda) \qquad \textit{Equation 71}$$

$$q_k \propto\ f(Y_i|\theta_k,\varsigma) \qquad \textit{Equation 72}$$

Where $f(Y_i / \theta_k.\varsigma)$ is the density of the marginal distribution of Y_i, q_o and q_k are standardised to sum to 1.

To carry out MCMC sampling we need to evaluate the complex integral q_o this can be difficult if G_o is non-conjugate which is the main drawback of the algorithm. But in our case the priors are conjugate as we can use normally distributed variables and adapt these for binary variables so that the binary variable is being driven by an underlying normally distributed variable. See chapter 5.10 later on for implementing the Escobar and west algorithm for the model. The method also came under criticism because the marginal distribution in some cases maybe sampled inefficiently thus taking a longer time to converge (Dey, Muller et al. 1998).

A new marginal method was introduced by (Bush and MacEachern 1996) this works by marginalising over both G and also θ_i To form a conditional distribution on a cluster membership variable S for each point rather than updating a mixture parameter, this also relies on conjugatecy although a newer method has been adapted to allow for non-conjugatecy (MacEachern and Muller 1998). Although this method is newer and highly suitable for situations where a Dirichlet process is needed over real data in order to obtain cluster membership. Problems occur when applying the Bush and MacEachern MCMC algorithm over a latent variable as by marginalising over the *G* and θ_i parameters we do not compute the actual Z_i of the latent variable *Z* as these cannot be inferred because the prior parameter for them has been integrated out and we need these to visualise the latent variables and to use them to infer in the model using the non-Dirichlet part of the model.

The Escobar and West algorithm can actually cancel down to a smaller number of clusters as the number of clusters is always smaller than the number of points so we get I+1 unique clusters in the data leading to the new equation 73

$$[\theta_i|\{\theta_k, k \neq i\}, Y, \varsigma, \lambda] \sim$$
$$q_o \, . \, G_{lim}(\theta_i|\lambda, Y_i, \varsigma) + \sum_{k=1}^{I} n^k q_k^* \, . \delta(\theta_i|\theta_k^*) \qquad \textit{Equation 73}$$

Where the θ_k^* are the cluster specific parameters and the n^k are the number of points in cluster k for that iteration.

An extra step can be added to infer on the θ_k to determine their value depending on the prior and the data points the mixture contains this is instead of arbitrarily selecting random values from the prior (Bush and MacEachern 1996), this improves efficiency of the algorithm to match the Bush algorithm, without losing the mixture parameters. This can be calculated as in equation 74.

$$p(\theta_k^*|Y, S, I, \varsigma, \lambda) \propto \prod_{i \in K_k} f_i(Y_i|\theta_k^*, \varsigma) \, dG_o(\theta_k^*, \lambda) \qquad \textit{Equation 74}$$

Where S is a membership indicator and K_k={I : S_i =k}

Another marginal method is that of (Jain and Neal 2004) who created a Split-merge algorithm this method used Metropolis-Hastings sampling procedure to determine whether a group of data should be split up in to two distinct groups or merged with another group, although this tackles the issue of possibly getting stuck in low probability solutions in the MCMC marginal techniques it has several disadvantages it is more computational expensive, it only allows the splitting up into two groups or

merging of two groups and although it allows lots of points to be moved in one iteration this is at the expensive of not allowing only one point to move at a time which could be the case if mixtures overlap. It also only out performs the Escobar Gibbs sampling method when there are two groups with similar parameters, with other cluster situations favouring the Gibbs sampling marginal methods above.

5.8.2 Conditional MCMC algorithms

These algorithms do not integrate out any parameters so inference can be obtained on all parameters if required. Due to the infinite nature of Dirichlet process mixtures the conditional algorithms are tricky to compute and can be harder to implement as they can contain an infinite amount of parameters and mixtures. The only algorithms found for full conditional techniques were a retrospective algorithm that assigns cluster/mixture membership first and then calculates a probability for cluster allocation (Papaspiliopoulos and Roberts 2008).

This allows the clusters to be identifiable thus combating issues associated with label switching of mixtures; this is when one mixture swaps labels with another mixture even though they still contain the same individuals in them. Even though there may be many mixtures and it also allows the probabilities of belonging to the infinite mixtures to be computed which are integrated out in the marginal methods. This retrospective Dirichlet process however is more complicated to code and is slower to converge when compared to the marginal methods of (Escobar and West 1995) and (MacEachern and Muller 1998) above for this reason it was not investigated further see (Papaspiliopoulos and Roberts 2008) for further details of the code.

5.8.3 Approximate Dirichlet process mixtures

The approximated Dirichlet process comes from the description of the Dirichlet process mixture being the infinite limit of a mixture model. The infinite mixture model, Dirichlet mixture model, can be approximated by a finite mixture model with a large number of mixtures (Ishwaran and James 2002).

Consider the Dirichlet process again as a discrete distribution over an infinite number of real points. The probabilities associated with these points can be constructed by a stick breaking process. Imagine a stick of unit length, we break a piece off the stick s_1 and assign it to be the probability p_1 of the point x_1, the remainder of the stick having magnitude $(1-s_1)$. Then break another piece of the rest of the stick s_2 to create the probability p_2 belonging to x_2, $p_2=(1-s_1).s_2$ and so on, the last remaining part could be infinitely small, an infinite amount of points are used (Ishwaran and James 2002).

$$G = \sum_{k=1}^{\infty} p_k \, N(\theta_k) \qquad \textit{Equation 75}$$

$$\theta_k \sim G_o \qquad \textit{Equation 76}$$

The approximate or truncated Dirichlet Process/Dirichlet Process Mixture model was suggested by (Ishwaran and Zarepour 2002) and has since been implemented in WinBUGS (Ohlssen, Sharples et al. 2007) for measured variables. The truncated Dirichlet Process Mixture is similar to a full Dirichlet Process mixture except that the maximum number of distributions (N) in G_o is fixed in advance. This model can be thought of as limiting the number of breaks in the stick breaking process to N-1, the last part of the stick being equal to 1 minus the other parts of the stick, equation 79.

$$\sum_{k=1}^{\infty} p_k \delta_{\theta_k} \approx \sum_{k=1}^{N} p_k \, \delta_{\theta_k} \qquad \textit{Equation 77}$$

$$\sum_{k=1}^{\infty} p_k \, N(\mu_k, V^2) \approx \sum_{k=1}^{N} p_k \, N(\mu_k, V^2) \qquad \textit{Equation 78}$$

$$and \; p_k = s_k \prod_{k<j} (1 - s_j) \qquad \textit{Equation 79}$$

$$where \; s_j \sim Beta(1, \alpha) \qquad \textit{Equation 80}$$

Although setting boundaries on the number of distributions is a limitation, the maximum number of groups allowed can be set to be much higher than is thought likely to occur. This is not as limiting as the k-means or finite mixture models in which the actual number of clusters has to be specified.

In fact as long as the number of mixtures fixed a prior N is larger than the actual number of mixtures needed so that the Dirichlet process mixture converges to, then the algorithm works in a very similar way then the full Dirichlet process. The truncated/approximate Dirichlet process mixture can be used to compute a very good approximation to the full Dirichlet process mixture which is faster than marginal methods (Ishwaran and Zarepour 2000), can be used with or without conjugatecy (Ishwaran and Zarepour 2000) and has been adapted for the powerful Bayesian software WinBUGS (Ohlssen, Sharples et al. 2007). This makes for a powerful argument to use truncated/approximate Dirichlet process mixtures but if possible it is better not to use approximations (Calla, 2008). This method is also conditional so

inferences can be made on the probability associated with the clusters we leave the truncated Dirichlet process here but will come back to it for chapter 10 where we use the truncated model to fit non-conjugate data for an application of the Dirichlet process to a clinical trial concerning cancer survival data.

5.9 Latent Dirichlet Process distributed variables

Latent variable distributions are usually assumed to be either normally distributed (Usually standard normally distributed). This was previously described in the chapter on factor analysis, chapter 4.4 but we now describe it in a Bayesian model notation with a latent variable Z rather than a factor F for the j variables and i individuals. As outlined in equations 81-84.

$$Y_{ij} \sim N(\theta_{ij}, \sigma^2{}_j) \qquad \textit{Equation 81}$$

$$Y_{ij} \sim N(\theta_{ij}, \sigma^2{}_j) \qquad \textit{Equation 82}$$

$$\theta_{ij} = \beta_{0j} + \beta_{1j}.Z_i \qquad \textit{Equation 83}$$

$$Z_i \sim N(0, 1) \qquad \textit{Equation 84}$$

Where Y_j are the j measured variables, θ_{ij} is the subject specific mean of the j variable, $\sigma^2{}_j$ is the variance of variable Y_j, β_{0j} and β_{1j} are the coefficients of the j variable and Z_i is the i score of the standard normally distributed latent variable.

This could be an over simplification of the latent variables distribution as it is impossible to determine the distribution of a latent variable from the data (Lee, Lu et al. 2008). Important variation information could be lost when the data does not conform to parametric distributions. The latent variable model can be adapted to include a latent variable that is distributed with a Dirichlet process normal mixture. This is described below for the case of one latent variable and j measured variables in statistical model terms in equation 85-90

$$Y_{ij} \sim N(\theta_{ij}, \sigma^2{}_j) \qquad \textit{Equation 85}$$

$$\theta_{ij} = \beta_{0j} + \beta_{1j}.Z_i \qquad \textit{Equation 86}$$

$$Z_i \sim D(\alpha, G_o) \qquad \textit{Equation 87}$$

$$\mu_i, V_i \sim N(\mu_0, V_0).IGamma(a_0, b_0) \qquad \textit{Equation 88}$$

$$G_o \sim N(\mu_i, V_i) \qquad \textit{Equation 89}$$

$$\alpha \sim Gamma(a, b) \qquad \textit{Equation 90}$$

Where Y_j are the j measured variables, θ_{ij} is the subject specific mean of the j variable, $\sigma^2{}_j$ is the variance of variable Y_j, β_{0j} and β_{1j} are the coefficients of the j variable, Z is a latent variable, α is the precision parameter of the Dirichlet process mixture with a gamma prior Gamma(a,b) , G_o is the location parameter of the Dirichlet process, μ_i and V_i are the means and variances of the Z_i with the μ_i and V_i having a composite normal inverse gamma prior with normal parameters μ_0 and V_0 and gamma parameters a_0 and b_0.

5.10 Conclusions on Implementing Dirichlet process normal mixtures over a latent variable

We now look at implementation of the Dirichlet process normal mixture latent variable model. The Escobar and West algorithm was used to code the Dirichlet process over the latent variable in a latent variable model as this is perfect for our needs as it takes advantage of the conjugatecy without removing the values of the latent variable so these can be computed, displayed and hypothesis tested on if required. It is also not an approximation and is quite fast to converge.

As the full Dirichlet process was not able to be carried out in WinBUGS (Lunn 2000) due to its infinite mixture nature the model was coded in R language (R Development Core Team 2009) . First the latent variable model was to be coded into R using only one normally distributed latent variable and four normally distributed variables.

$$Y_{ij} \sim N(\theta_{ij}, \sigma^2{}_j) \qquad \textit{Equation 91}$$

$$\theta_{ij} = \beta_{0j} + \beta_{1j}.Z_i \qquad \textit{Equation 92}$$

$$Z_i \sim N(0,1) \qquad \textit{Equation 93}$$

$$\beta_{0j} \sim N(0{,}100) \qquad \textit{Equation 94}$$

$$\beta_{1j} \sim N(0{,}100) \qquad \textit{Equation 95}$$

$$1/\sigma^2 \sim Gamma(0.01{,}0.01) \qquad \textit{Equation 96}$$

Where Y_j are the j measured variables j=1 ,2, 3 and 4, θ_{ij} is the subject specific mean of the j variable , $1/\sigma^2$ is the precision of variable Y_j , β_{0j} and β_{1j} are the coefficients of the j variable and Z_i is the i score of the standard normally distributed latent variable.

The posterior distributions were derived by hand by multiplying the prior and likelihood distribution together and rearranging to find appropriate distributions. These were then coded for in R and then tested using simulated data to determine errors in the code and to correct for these.

Once the normally distributed latent variable model was established and error free. The coding then turned towards the Dirichlet process Normal mixture code. To start a simple version of the Dirichlet process normal mixture was coded only over one manifest variable and with no hyper parameters on the Dirichlet process parameters. Again this was checked for errors and limitations of the model. The model coded for is described below.

$$Y_i \sim D(\alpha, G_o) \qquad \textit{Equation 97}$$

$$G_o \sim N(\mu_i, V_i) \qquad \textit{Equation 98}$$

$$\mu_i \sim N(0, V_i) \qquad \textit{Equation 99}$$

$$1/V_i \sim Gamma(0.01, 0.01) \qquad \textit{Equation 100}$$

$$\alpha = 1 \qquad \textit{Equation 101}$$

This model worked well but it was possible that the priors on the Dirichlet parameters may be too constrained for some situations. So hyper priors were added to the mean parameter μ_i and the Dirichlet precision parameter α giving a new less restricted model. In order to calculate the integrals involved in the MCMC sampling a trick was needed for the specification of the hyper parameters this was letting the variance of μ_i be related to the variance by a scalar τ which also has a prior associated with it , see equation 107 (Escobar and West 1995).

$$Y_i \sim D(\alpha, G_o) \qquad \textit{Equation 102}$$

$$G_o \sim N(\mu_i, V_i) \qquad \textit{Equation 103}$$

$$\mu_i \sim N(m, \tau V_i) \qquad \textit{Equation 104}$$

$$V_i^{-1} \sim Gamma\left(\frac{s}{2}, \frac{S}{2}\right) \qquad \textit{Equation 105}$$

$$\alpha \sim Gamma(2, 4) \qquad \textit{Equation 106}$$

$$\tau^{-1} \sim Gamma(0.1, 0.1) \qquad \textit{Equation 107}$$

$$m \sim N(0, 1000) \qquad \textit{Equation 108}$$

The reason the parameter τ was needed when adding hyper parameters to the Escobar and west MCMC techniques can be seen by investigating the MCMC algorithms for the parameters involved taking the Escobar and west model from before, where the value

θ_i in the Escobar and west algorithm represents the set of two parameters for a normal distribution mixture, the mean μ_i and the variance V_i of the Y_i's .

$$[\theta_i|\{\theta_k, k \neq i\}, Y, \varsigma, \lambda] \sim$$
$$q_o \, . \; G_{lim}(\theta_i|\lambda, Y_i, \varsigma) + \sum_{k=1}^{I} n^k q_k^* \, . \delta(\theta_i|\theta_k^*) \qquad \textit{Equation 109}$$

$$q_o \propto \alpha \int f(Y_i|\theta_i, \varsigma_i) \; . dG_o(\theta_i|\lambda) \qquad \textit{Equation 110}$$

$$q_k \propto \; f(Y_i|\theta_k, \varsigma_i) \qquad \textit{Equation 111}$$

To obtain q_o we have to use the special prior formulation as stated in equation 112 and 113 (Escobar and West 1995).

$$\mu_i \sim N(m, \tau V_i) \qquad \textit{Equation 112}$$

$$V_i^{-1} \sim Gamma\left(\frac{s}{2}, \frac{S}{2}\right) \qquad \textit{Equation 113}$$

Where m is the mean of the prior of μ_i , τ is the scale factor of the variance, s/2 is the shape of the gamma distribution for V^{-1} and S/2 is the rate. By using these formulae it allows the MCMC algorithms of parameters to cancel out to create common distributions which speed up the algorithm for the posterior probability for the parameters below see equations 114-116.

$$(\mu_i|V_i) \sim N\left(\frac{(m + \tau . Y_i)}{(1 + \tau)}, \frac{\tau V_i}{1 + \tau}\right) \qquad \textit{Equation 114}$$

$$V^{-1} \sim Gamma(\frac{1+s}{2}, \frac{S + \frac{(Y_i - m)^2}{1+\tau}}{2}) \quad \textit{Equation 115}$$

$$q_k \propto \frac{e^{\frac{-(Y_i-\mu_i)^2}{2.V_k}}}{(2.V_k)^{\left(\frac{1}{2}\right)}} \quad \textit{Equation 116}$$

This allows the difficult q_o integral to be computed thusly

$$q_o \propto \alpha \int f(Y_i|\theta_i, \varsigma) \ . dG_o(\theta_i|\lambda) \quad \textit{Equation 117}$$

In our code f is normally distributed below as θ equals the two parameters μ V and ς=0 as at present as there are no other non Dirichlet process parameters in this model.

$$f(Y_i|\theta_i, \varsigma_i) = N(Y_i|(\mu_i, V_i) \quad \textit{Equation 118}$$

Remembering previously

$$G_o \sim N(\mu_i, V_i) \quad \textit{Equation 119}$$

$$\mu_i \sim N(m, \tau V_i) \quad \textit{Equation 120}$$

$$V_i^{-1} \sim Gamma\left(\frac{s}{2}, \frac{S}{2}\right) \quad \textit{Equation 121}$$

Stating the generic normal and Gaussian distribution formulae in equations 122 and 123.

$$N(Y|\mu,V^{-1})\sim\frac{V^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}.e^{-\frac{1}{2}V^{-1}(y-\mu)^2} \qquad \textit{Equation 122}$$

$$Gamma(V|a,b)\sim\frac{S^{\frac{k}{2}}.V^{\frac{(k-1)}{2}}.e^{-\frac{1}{2}(SV)}}{2^{\frac{k}{2}}.\Gamma\,(k/2)} \qquad \textit{Equation 123}$$

We obtain equations 124-129

$$f(Y_i|\theta_i,\varsigma_i)=\frac{{V_i}^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}.e^{-\frac{1}{2}{V_i}^{-1}(y_i-\mu_i)^2} \qquad \textit{Equation 124}$$

$$G_0(\theta_i|\lambda)=N(m,\tau V_i).Gamma(\frac{s}{2},\frac{S}{2}) \qquad \textit{Equation 125}$$

$$dG_0(\theta_i|\lambda)=\ N(m,\tau V_i).Gamma\left(\frac{s}{2},\frac{S}{2}\right)d\mu_i.dV_i \qquad \textit{Equation 126}$$

$$dG_0(\theta_i|\lambda)=\left(\frac{\tau{V_i}^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}.e^{-\frac{1}{2}\tau{V_i}^{-1}(\mu_i-m)^2}\right).\left(\frac{S^{\frac{k}{2}}.V_i^{\frac{(k-1)}{2}}.e^{-\frac{1}{2}(SV_i)}}{2^{\frac{k}{2}}.\Gamma\left(\frac{k}{2}\right)}\right)d\mu_i.dV_i$$

Equation 127

$$q_o\propto\alpha\int f(Y_i|\theta_i,\varsigma_i)\ .dG_o(\theta_i|\lambda) \qquad \textit{Equation 128}$$

$$q_o \propto \alpha \int\int \left(\frac{V_i^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} . e^{-\frac{1}{2} V_i^{-1}(y-\mu_i)^2} \right) . \left(\frac{\tau V_i^{-\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} . e^{-\frac{1}{2}\tau V_i^{-1}(\mu_i - m)^2} \right) . \left(\frac{S^{\frac{k}{2}} . V_i^{\frac{(k-1)}{2}} . e^{-\frac{1}{2}(SV_i)}}{2^{\frac{k}{2}} . \Gamma\left(\frac{k}{2}\right)} \right) d\mu_i . dV_i$$

Equation 129

By rearranging the formula and solving the integral we obtain equation 130

$$q_o \propto \alpha . c(s) . \frac{\left[1 + \frac{(Y_i - m)^2}{sM}\right]^{-\frac{1+s}{2}}}{M^{\frac{1}{2}}}$$

Equation 130

Where

$$c(s) = \Gamma\left(\frac{1+s}{2}\right) . \Gamma(s/2)^{-1} s^{-1/2}$$

Equation 131

$$M = \frac{(1+\tau)S}{s}$$

Equation 132

So we can now carry out the iterative steps needed in the MCMC algorithms as the value of q_o has been evaluated for a Dirichlet process with hyper-parameters over one measurable variable.

The two state model, latent variable model, equation 133-137 and the Dirichlet process mixture model, equation 138-142 were then combined together to create the latent variable Dirichlet process normal mixture model, see below.

$$Y_{ij} \sim N(\theta_{ij}, \sigma^2{}_j) \qquad \textit{Equation 133}$$

$$\theta_{ij} = \beta_{0j} + \beta_{1j}.Z_i \qquad \textit{Equation 134}$$

$$\beta_{0j} \sim N(0,100) \qquad \textit{Equation 135}$$

$$\beta_{1j} \sim N(0,100) \qquad \textit{Equation 136}$$

$$1/\sigma^2 \sim Gamma(0.01,0.01) \qquad \textit{Equation 137}$$

$$Z_i \sim D(\alpha, G_o) \qquad \textit{Equation 138}$$

$$G_o \sim N(\mu_i, V_i) \qquad \textit{Equation 139}$$

$$\mu_i \sim N(m, \tau V_i) \qquad \textit{Equation 140}$$

$$V_i{}^{-1} \sim Gamma\left(\frac{s}{2}, \frac{S}{2}\right) \qquad \textit{Equation 141}$$

$$\alpha \sim Gamma(2,4) \qquad \textit{Equation 142}$$

In this part of the code it was first discovered about the many equivalent solutions the latent variable model had this was due to the latent variable not having specified mean and variance, as previously in the latent variable model these were kept as a mean of 0 and a standard deviation of 1. Thus constraints were needed to find a unique solution

that was also equivalent to all the other solutions this was done by using the first variable as a factor anchor by keeping the latent variable equation parameters constant for this variable. This is equivalent to keeping β_{01}=0 and β_{11}=1. This is often done in factor analysis and can be seen as a methodology in chapter 4.3. This allowed a single solution to be found that was equivalent to the other solutions. Although this solution was found the model took a long time, several days, to converge for this reason the Escobar and West algorithm was adapted to include the added parameter mixture conditioning step to speed up convergence, this was basically to determine the mean and variance of the individual infinite mixtures by conditioning on them so that there are not sampled randomly.

Conditioning on the μ_k and V_k we obtain the new MCMC algorithms for sampling μ_k and V_k , equation 143 and 144, depending on the individual latent variable score.

$$\mu_k^* \sim N\left(\frac{m.\tau + \sum_{i \epsilon K_k} Z_i}{n^k}, \ \frac{n^k + \tau}{V_k^*} \right) \qquad \textit{Equation 143}$$

$$V_k^* \sim I.G\left(\frac{s + n^k + \tau + 1}{2} \ , \frac{S + \tau(\mu_k^* - m)^2 + \sum_{i \epsilon K_k} (Z_i - \mu_k^*)^2}{2} \right) \qquad \textit{Equation 144}$$

The MCMC sampling methodology can now be summarised thus

Step 1 Choose starting values for all parameters, both parametric and Dirichlet process mixture

Step 2 Gibbs sample from the posterior equations of the non-Dirichlet process mixture parameters the ones from the latent variable part of the model.

Step 3 Sample the values of the latent variable Z_i using the non-Dirichlet process values and the Dirichlet values θ_i i.e. μ_i and V_i

Step 4 Sample using the $\theta_i|\{\theta_k^*, k\neq i\}, Z_i, \varsigma, \lambda\}$ for all of the n points i to

Step 5 Condition on the θ_k, using cluster membership of the n points, i.e. find the new means and variance of the mixtures with the new points contained in them.

Step 6 Condition on the hyper parameters of the Dirichlet process m, τ and α

Step 5 Repeat from Step 2-6 with the new sampled values until convergence occurs.

The Dirichlet process normal mixture latent variable model now worked well but could be very slow when programmed into R, taking days to converge. This is due to the compiled nature of R and the sampling design of the Dirichlet process, updating each point in the latent variable separately for each iteration.

As it was the Dirichlet process sampling part of the code that was taking the longest, the part containing the Dirichlet process mixture was taken out of the R language and adapted and programmed and compiled for use in R, but in C language as a C function, this was done to speed up the process as C language although basic is very fast.

The C program containing the Dirichlet process part of the model was called by R and ran. The idea of reprogramming Dirichlet process in a faster language can also be found in the literature (Hoff 2005) and in the Dirichlet process R package, DPPackage (Jara, Garcia-Zattera et al. 2005). The C function I created containing the Dirichlet part

of the code was then downloaded into R and ran in R; this reduced the time taken to converge to hours rather than days. This allowed the Dirichlet process mixture model to be implemented and allowed to converge in a realistic time frame. Allowing results to be obtained overnight. First the C language had to be understood and tested using simple examples and then functions were coded again using simple examples so that they could be called from R and after familiarity with C had been established. The Dirichlet sampling part was coding in C and made available in R for LINUX. Allowing summary statistics to be obtained and simulations ran in a shorter time period. To see the code for the C program, see appendix 1 and for the full R code for the Dirichlet process for 9 variables see appendix 2.

5.11 Chapter closing statement

We have described the theory on how to implement a Dirichlet process and we have used that theory to combine the Dirichlet process normal mixture model to a latent variable in a latent variable model. This involved making suggestions on which algorithm to use, how to code the model into available statistical programming software how to adapt the two models so they can be merged and how to speed up the algorithm in order to achieve results in a realistic time frame we now look to test the model more formally by carrying out simulations in the next two chapters, chapter 6 and 7, a range of different clustering scenarios were used to determine if the model can detect these different cluster patterns and if we can determine whether the clusters are down to sampling or reflect real clusters

Chapter 6. Simulation Using One Latent Variable and Differing Mixtures

6.1 Chapter overview

Different cluster/mixture patterns are simulated and the Dirichlet process normal mixture latent variable model tested to determine if these mixture patterns can be detected by the model and if correct cluster membership can be returned. If the correct number of clusters are not returned then reasons are found in the simulation. The simulations are carried out for 200 individuals and then 500 individuals for continuous data to show consistency. Binary variables were also simulated to determine if these could be used with the continuous data to determine the underlying latent structure.

6.2 Introduction

In order to check whether the models suggested in the Dirichlet process mixture chapter can detect the underlying structure of a latent variable with 6 normally distributed manifest variables data was simulated for 10 different latent variable distribution scenarios. These same scenarios where then used again to simulate 4 normally distributed variables and 2 binary ones. To test the impact of combining binary variables with normally distributed variables in order to make recommendations relevant to the severe asthma data.

There are infinitely many possible latent variable distribution scenarios so I chose the 10 different scenarios that possibly explained the severe asthma data variation discovered from the literature (see chapter 2 on severe asthma phenotypes). The

latent variable was often chosen to be represented by two mixtures, reflecting the most common sub-groups seen in the literature that of atopic or non- atopic, or eosinophilic/ non-eosinophilic. The latent variable model was fitted with various differing examples of sub-groups, i.e. same size groups, one small group and one big group, and differing distances between the groups. Each scenario was simulated 10 times each with a different seed for the random sampling of the latent variable. The normal variables were calculated from the latent variables using constants to linear transform the latent variables into measured ones using a different standard deviation for each one.

For binary variables linear transformations were again used to transform the latent variable into a standard distributed latent variable. The standard latent variable was then used to simulate a binary variable with values of 1 if the distribution was larger than 0 and a binary value of 0 if the standard distributed value was less than 0.

Once the Bayesian models were ran the models were tested for convergence this was done using the Heidelberger diagnostic criteria, (Heidelberger and Welch 1981) for testing non-convergence of parameters in a Bayesian MCMC model and by viewing the density and trace plots of parameters(Gelman, Carlin et al. 2004). Statistics were taken from the iterations to determine

- If the mixtures could imply a multi modal distribution,

  This was derived from a frequentist hypothesis test for multi modality of static data called a diptest, to obtain a mean diptest for the latent variable,(Hartigan and Hartigan 1985).
- Whether the correct number of clusters could be obtained from the model,

Using hierarchical clusters of the probability of not belonging in a mixture with another subject(Medvedovic and Sivaganesan 2002).

- Whether the correct cluster membership could be obtained,

  This was defined by cutting the dendrogram obtained from the probability clustering above at the point of maximum difference between clusters(Everitt 2001).

6.3 Generating simulations

Variables are generated using the following mixture methodology to create a latent variable distribution.

$$Z_i \sim \pi_k . N(\theta_k , {\sigma_k}^2) \qquad \textit{Equation 145}$$

$$\sum_{k=1}^{M} \pi_k = 1 \qquad \textit{Equation 146}$$

Where M is the number of mixtures and π_k is the proportion of the *k* normal distribution in the latent variable Z, and $N(\theta_k , \sigma_k)$is a normal distribution with mean θ_k and variance ${\sigma_k}^2$.

Normal variables are generated using the latent variable above in the formula below.

$$Y_{ij} \sim N\left(\theta_j, {\sigma_j}^2\right) \qquad \textit{Equation 147}$$

$$\theta_j = \beta_{0j} + \beta_{1j} . Z_i \qquad \textit{Equation 148}$$

Where, β_{oj}, β_{1j}, σ_j are held constant over the simulations and scenarios, Y_{ij} is a measured variable for subject i and variable j distributed by $N(\theta_j, \sigma_j)$, a normal distribution with mean θ_j and standard deviation σ_j. Values of β_{oj}, β_{1j}, σ_j were chosen to represent a mixed variety of possible variables on differing scales and magnitudes.

The binary variables were again generated from the latent variable this time truncating a normal variable derived from the latent variable to obtain either a 1 or a 0 using the formula below

$$Y^B{}_{ij} = 1 \;\; if \;\; Y^*{}_{ij} > 0 \qquad \textit{Equation 149}$$

$$Y^B{}_{ij} = 0 \;\; if \;\; Y^*{}_{ij} < 0 \qquad \textit{Equation 150}$$

$$Y^*{}_{ij} \sim N(\theta_j, 1) \qquad \textit{Equation 151}$$

$$\theta_j = \beta_{0j} + \beta_{1j}.Z_i \qquad \textit{Equation 152}$$

Where β_{oj}, β_{1j}, σ_j are arbitrary selected integers held constant over the simulations and scenarios, $Y^B{}_{ij}$ is a binary variable for subject i and variable j distributed by a truncated normal distribution $N(\theta_i, 1)$ a normal distribution with mean θ_j and standard deviation 1, taking values $Y^*{}_{ij}$. For values of β_{oj}, β_{1j}, σ_j for each of the 6 variables for simulating the 6 normally distributed variables see table 6.3.1 and for the simulations using 4 normally distributed variables and 2 binary variables see table 6.3.2 these were chosen to represent a mixed variety of possible variables on differing scales and magnitudes.

Table 6.3.1 Parameters chosen for simulations in latent variable model for 6 normally distributed variables

Variable Number Y_j	β_{0j}	β_{1j}	σ_j
Y_1	5	3	1
Y_2	2	1	3
Y_3	2	2	1
Y_4	25	2	2
Y_5	4	3	2
Y_6	100	1	2

Table 6.3.2 Parameters chosen for simulations in latent variable model for 4 normally distributed variables and 2 Binary distributed variables

Variable Number Y_j	β_{0j}	β_{1j}	σ_j
Y_1	5	3	1
Y_2	2	1	3
Y_3	2	2	1
Y_4	25	2	2
Y_5 Binary	0	2	1
Y_6 Binary	2	1	1

6.4 Dirichlet process normal mixture latent variable model (DPNMLVM)

For Normal Variables

$$Y_{ij} \sim N\left(\theta_{ij}, {\sigma_j}^2\right) \qquad \textit{Equation 153}$$

$$\theta_{ij} = \beta_{0j} + \beta_{1j}.Z_i$$ *Equation 154*

$$Z_i \sim D(\alpha, G_o)$$ *Equation 155*

$$G_o \sim N(\mu_i, V_i)$$ *Equation 156*

$$\mu_i \sim N(m, \tau V_i)$$ *Equation 157*

$$V_i^{-1} \sim G(0.01,0.01)$$ *Equation 158*

$$\alpha \sim G(2,4)$$ *Equation 159*

$$\tau^{-1} \sim G(1,1)$$ *Equation 160*

$$m \sim N(0,100)$$ *Equation 161*

For Binary Variables

$$Y^B{}_{ij} = 1 \ \ if \ \ Y^*{}_{ij} > 0 \qquad \textit{Equation 162}$$

$$Y^B{}_{ij} = 0 \ \ if \ \ Y^*{}_{ij} < 0 \qquad \textit{Equation 163}$$

$$Y^*{}_{ij} \sim N(\theta_{ij}, 1) \qquad \textit{Equation 164}$$

$$\theta_{ij} = \beta_{0j} + \beta_{1j}.Z_i \qquad \textit{Equation 165}$$

$$Z_i \sim D(\alpha, G_o) \qquad \textit{Equation 166}$$

$$G_o \sim N(\mu_i, V_i) \qquad \textit{Equation 167}$$

$$\mu_i \sim N(m, \tau V_i) \qquad \textit{Equation 168}$$

$$V_i{}^{-1} \sim G(0.01, 0.01) \qquad \textit{Equation 169}$$

$$\alpha \sim G(2, 4) \qquad \textit{Equation 170}$$

$$\tau^{-1} \sim G(1,1) \qquad \textit{Equation 171}$$

$$m \sim N(0,100) \qquad \textit{Equation 172}$$

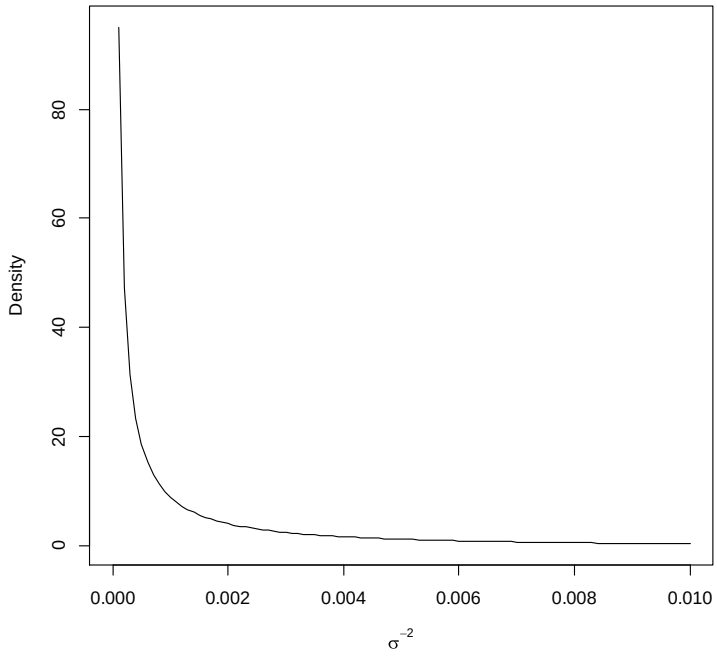
Where Y_{ij} represents the i individual of the j normally distributed variables θ_{ij} represent the mean of the Y_{ij} and σ_j the variance of the Y_{ij} variables β_{0j}, β_{1j} parameters of the regression of θ_{ij} on latent variable Z_i, $D(\alpha,G_0)$ is the Dirichlet process mixture with precision parameter α and centring distribution G_0, where G_0 is normally distributed with mean μ_i and variance V_i , Y^B_{ij} is the Binary variable of the i subject of the j binary variables which are distributed with a truncated Normal distribution $Y^*_{ij.}$

6.5 Priors

We now state the priors used in modelling for all the simulations. σ^2 is the variance parameter which was given a non informative inverse gamma prior with shape 0.01 and scale 0.01. The precision ($1/\sigma^2$) is given a non-informative prior over small values below 1 relating to a large range of variances see figure 6.5.1

$$\sigma^2{}_{ij} \sim I.G(0.01, 0.01) \qquad \textit{Equation 173}$$

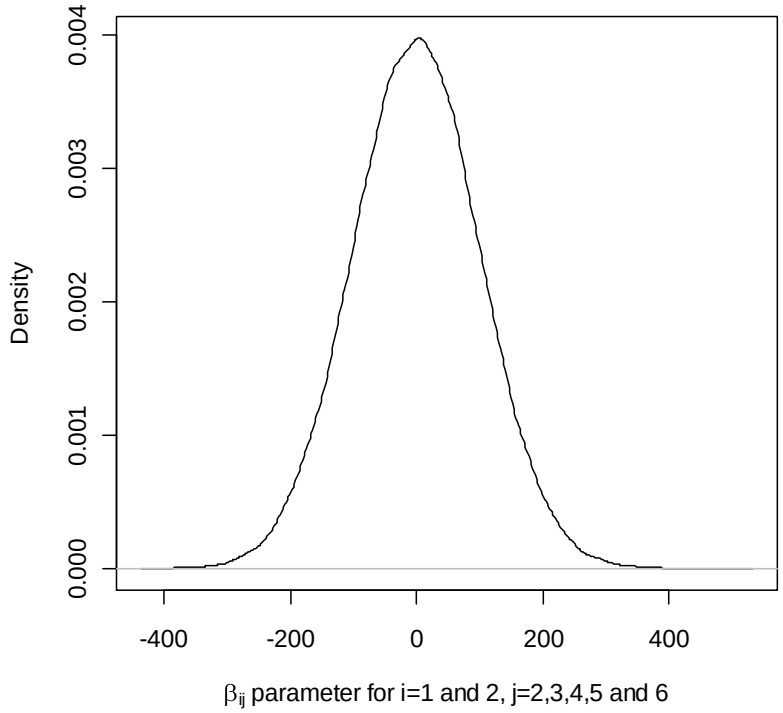
Figure 6.5.1 Graph of prior distribution used for σ^2 parameter in the model



Both priors for the β_{0j} and the β_{1j} for j = 2,…6 were normally distributed with standard deviation of 100 and mean 0 to achieve a very non-informative prior, see figure 6.5.2. β_{01} and the β_{11} were kept constant for identification purposes mentioned previously in chapter 5.10.

$$\beta_{0j} \sim N(0, 100) \qquad \textit{Equation 174}$$

Figure 6.5.2 Graph of prior distribution used for σ^2 parameter in the model



The precision parameter α of the Dirichlet process is given a gamma prior that looks informative but is actually not.

$$\alpha \sim G(1,2) \qquad \textit{Equation 175}$$

The prior has shape 1 and scale 2 these parameters cover values from 0 to 15 see table 6.5.3 and favour numbers of subgroups corresponding to a possible range of 1 to 40 mixtures, see graph 6.5.4 where n=200 using the approximation below {Escobar, 1994}.

$$E(k|\alpha,n) \approx \alpha.\ln(1+\frac{n}{\alpha}) \qquad \textit{Equation 176}$$

Figure 6.5.3 Graph of prior distribution used for α parameter in the model

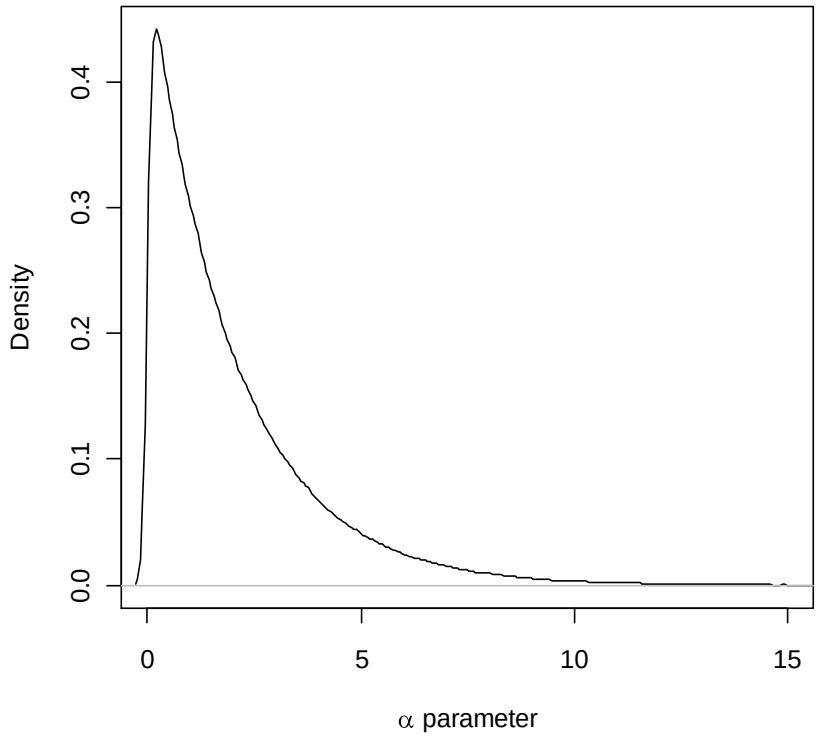
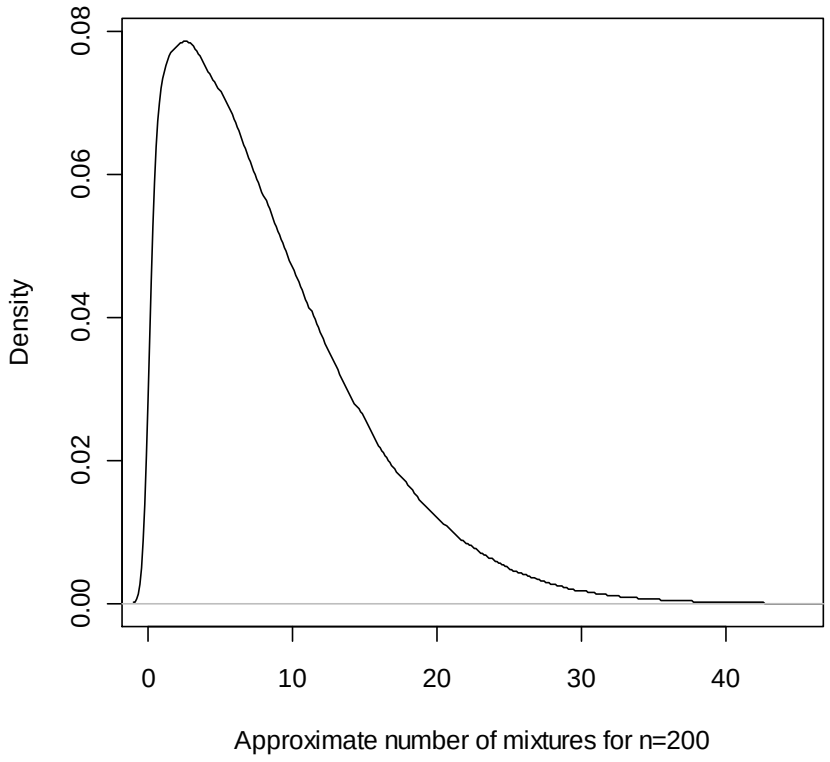


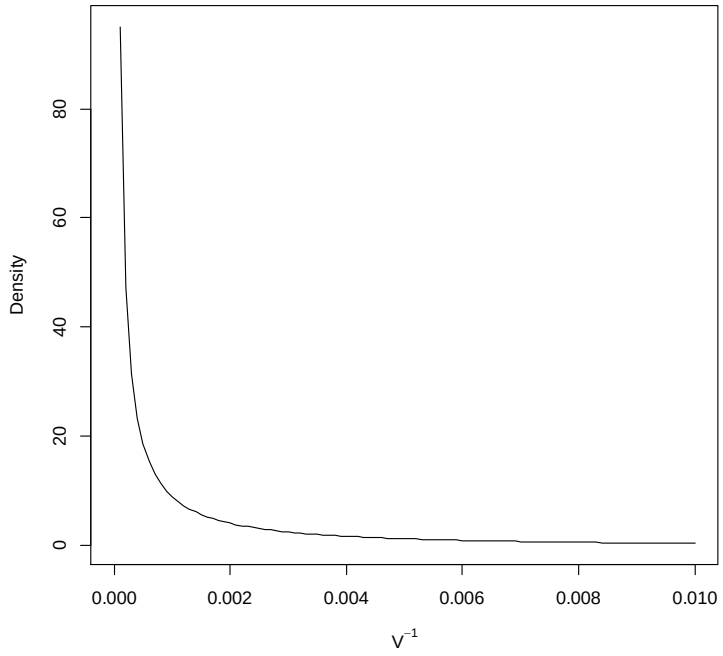
Figure 6.5.4 Graph of prior distribution used for k parameter in the model



For the prior for the variance of the mixture components the inverse of the variance was given a gamma prior with shape=0.01 and scale=0.01 again allowing an uninformative prior see figure 6.5.5

$$V_i^{-1} \sim G(0.01, 0.01) \qquad \textit{Equation 177}$$

Figure 6.5.5 Graph of prior distribution used for the V parameter in the model, the variance of a mixture.

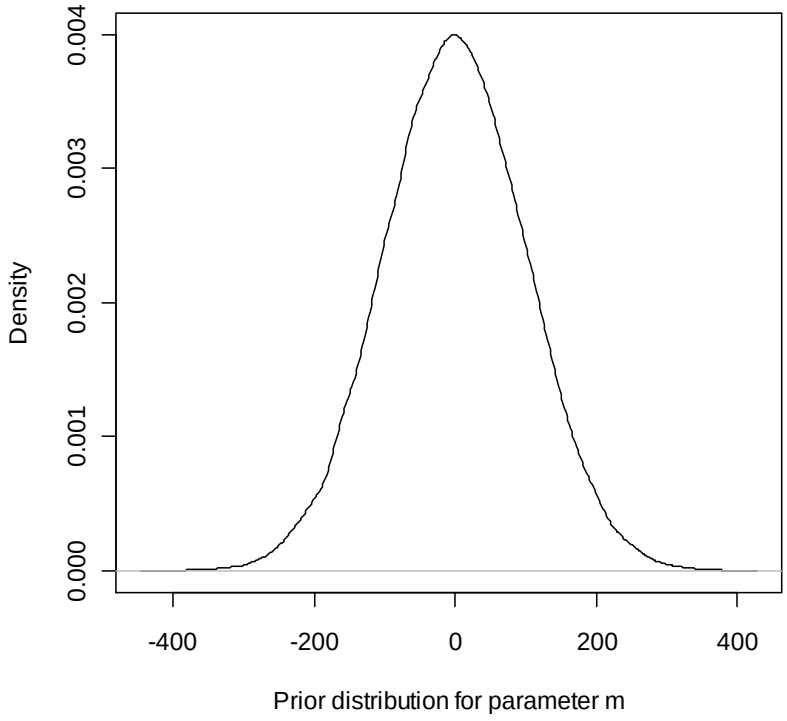


The m parameter is also given a non informative prior of mean 0 and standard deviation of 100 indicating no prior information on the parameter.

$$\theta_i \sim N(m, \tau . V_i)$$ *Equation 178*

$$m \sim N(0,100)$$ *Equation 179*

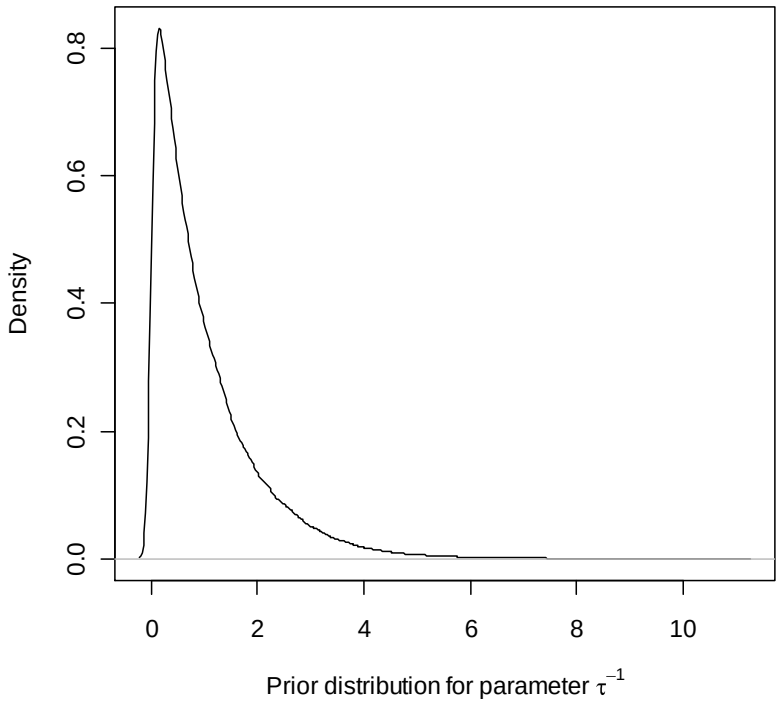
Figure 6.5.6 Graph of prior distribution used for the m parameter in the model the value of the mean of the mixture component



The tau parameter is given a vague inverse gamma prior with shape=1 and scale=1 to indicate that the τ^{-1} is usually between 0 and 8 corresponding to a large range of values of τ.

$$\tau^{-1} \sim G(1,1)$$ *Equation 180*

Figure 6.5.7 Graph of prior distribution used for the τ^{-1} parameter in the model the value of the mean of the mixture component.



6.6 Result determination

6.6.1 Convergence

Convergence was checked for the parameters using the Heidelberger and Welch's convergence diagnostic (Heidelberger and Welch 1981) the function used is in R, in the coda package (Plummer,2006). The Plummer version of the Heidelberger diagnostic returns a 1 for each time a parameter passes the convergence test and a 0 if a parameter fails. This test was chosen as it returns a simple yes or no answer to convergence and this is needed when testing a large number of parameters in Bayesian models as examining trace and density plots for 220 parameters for 200 subjects and 520 parameters for 500 subjects for each simulation and scenario would be impractical.

The test uses the null hypothesis that the values from the posterior of a parameter come from a stable distribution. It uses the Cramer-von-Mises statistic (T) to test the null hypothesis using the formula below.

$$T = \frac{1}{12n} + \sum_{i=1}^{n} [\frac{2i-1}{2n} - F(x_i)]^2 \qquad \textit{Equation 181}$$

Where n is the number of iterations x_1, x_2, …, x_n are draws from the posterior parameters distribution. F is the distribution of the x_i's assumed to be normally distributed for each parameter.

The value of T is checked against tabulated values to check for convergence, the larger the values of T the more likely we are to reject the null hypothesis. We use the value of T at a p-value of 0.05 as the cut off for accepting the null hypothesis.

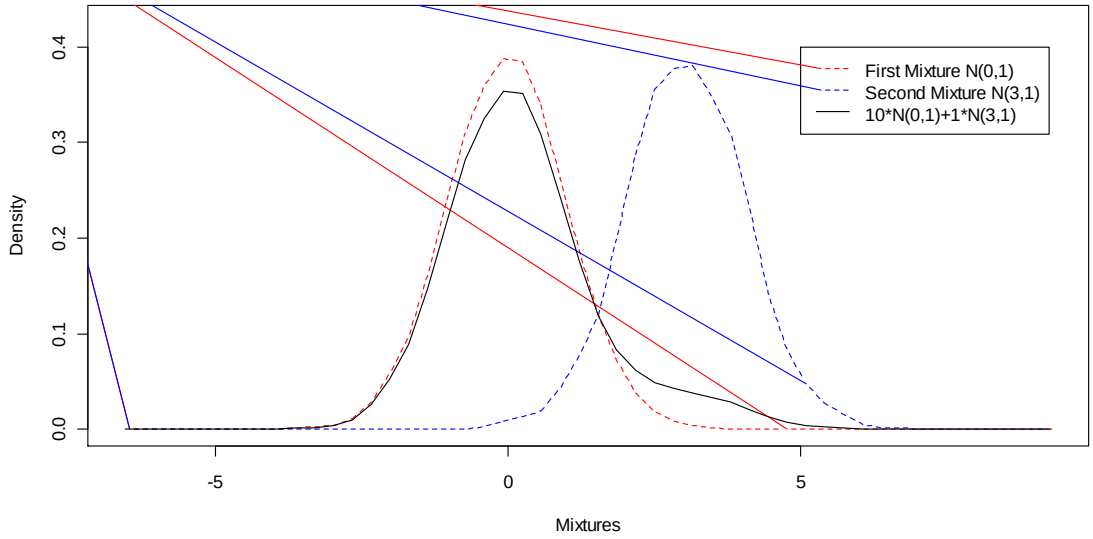
The result of convergence (YES/NO) of the parameters is expressed in each scenario simulation as a percentage of total parameters passing the test. If any of the parameters have not passed the convergence test this does not mean that convergence has not been achieved but rather the iterations for that parameter have not passed the strict condition $p \geq 0.05$ needed to pass the Heidelberger test of convergence.

For the parameters that failed the Heidelberg test a visual inspection was carried out using trace and density plots to check that convergence had actually been achieved. Convergence was achieved using 200,000 iterations and a burn-in of 150,000 iterations for all simulations.

6.6.2 Multimodality

A distribution can be made up of a number of mixtures when these are formed into the best fitting arrangement of mixtures and stated as a sub-grouping of the data we call these clusters. A distribution can be made up of many mixtures but can be described best using only a few clusters. Before determining the best number of clusters within a latent variable distribution we first need reasonable cause to assume there are any clusters. This is achieved by obtaining an indication of the clusters nature other than their size i.e. whether the clusters are separated or overlapping. If they are overlapping or very close together the clusters found in the statistical model may not represent true sub-groups but rather non-normal distributions approximated by clusters/mixtures of normal distributions. Typically these could be skewed normal distributions or heavy tailed normal distributions, see figure 6.6.2.1.

Figure 6.6.2.1. The graphs below seem to suggest that a mixture of 2 normal distribtuions can look like a skew distribution without obvious sub-groups.



In these incidences it is hard to determine the difference between mixtures that are true sub-groups and mixtures that approximate non–normal distributions. To clarify the nature of the mixtures a statistic derived from the dip test for multi-modality (Hartigan and Hartigan 1985) was obtained for the latent variable posterior distribution based on the iterations of the MCMC model.

Hartigan's dip test is predominantly used for hypothesis testing to determine if measured data is multi modal or unimodal. The dip test can be defined as the maximum difference, over all the sample points of the data, between the empirical distribution function and the unimodal distribution function that minimizes that maximum difference. This measure is best illustrated with graphs to demonstrate what the dip statistic actually measures. Using a distribution comprising of a normal distribution with mean 0 and standard deviation 1 we can see that the difference between the best fitting unimodal distribution and the empirical distribution is small, see figure 6.6.2.2. But for a multimodal distribution there is a bigger maximum difference between the best-fitting unimodal distribution and the empirical distribution, see fig 6.6.2.3.

Figure 6.6.2.2 for Z distribution consisting of 200 points generated from formula below, dip statistic=0.022 and is indicated in red, empirical distribution is in blue and the unimodal distribution is in black

$$Z \sim N\ (0,1) \qquad \textit{Equation 182}$$

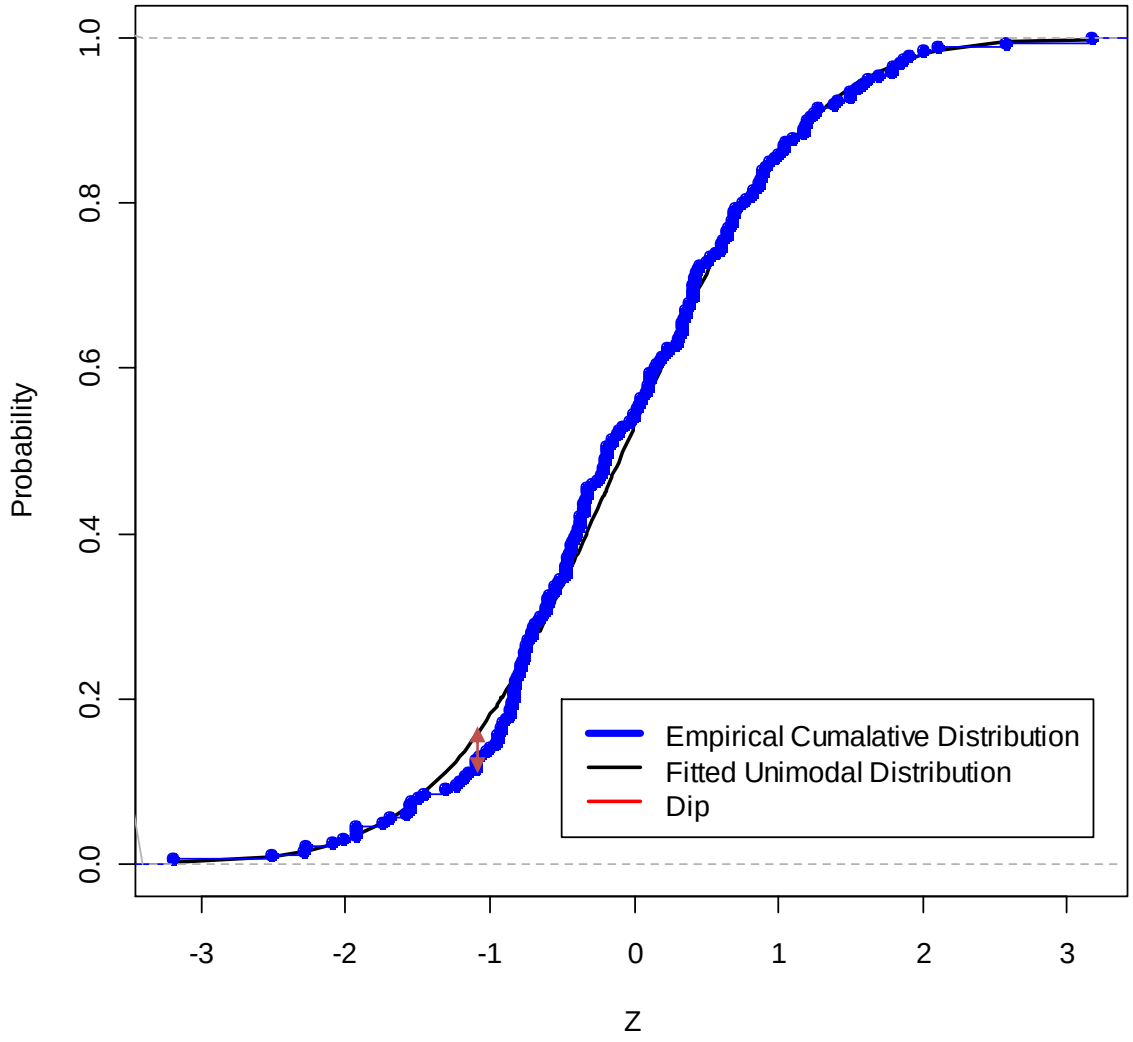
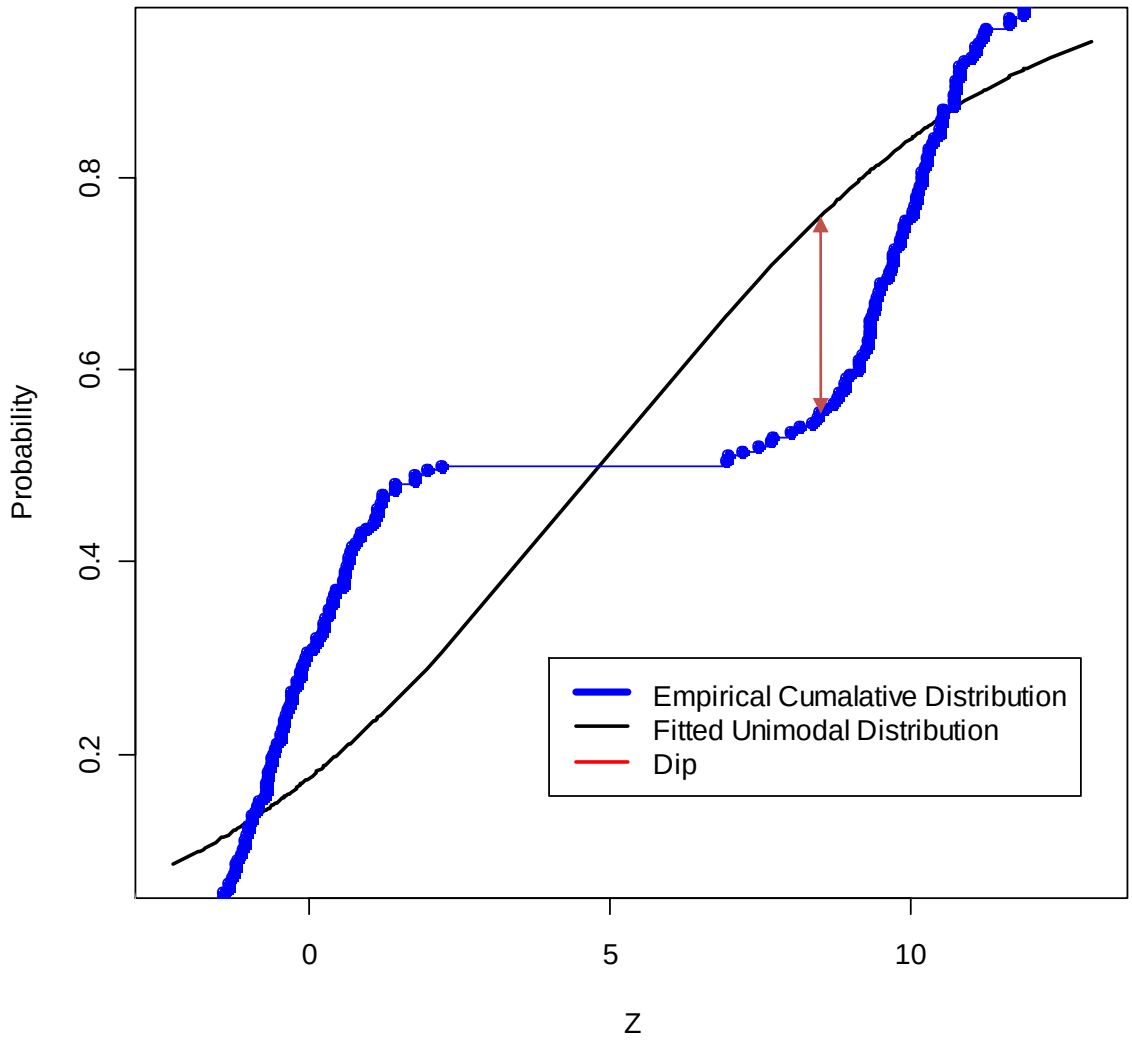


Figure 6.6.2.3 For Z distribution consisting of 200 points generated from formula below, dip statistic=0.138 and is indicated in red, empirical distribution is in blue and unimodal distribution is in black.

$$Z \sim 0.5.N\ (0,1) + \ 0.5.N\ (10,1) \qquad \textit{Equation 183}$$



The dip test assumes a null hypothesis that a distribution is unimodal and returns a statistic that can be compared in a table (see appendix A) to obtain a p value. This is fine for a single distribution to determine a p value that states whether the data

distribution is multimodal, but for distributions that change for each iteration as is in a Bayesian MCMC technique we need to summarise all of the dip statistics for all the iterations and compare these with the table to see if we can make inferences.

To summarise the many dip tests for all iterations, the mean dip statistic was tabulated along with the percentage of iterations that were found to be multimodal at p=0.05, dip statistic > 0.0185 for 200 subjects and dip statistic >0.0119 for 500 subjects (see Appendix 3 for dip statistic/p-value table).

6.6.3 Determination of number of clusters in the mixture model

The number of clusters was obtained by determining the probability of each point, Z_i belonging in the same mixture as every other point Z_j, $j \neq i$. This was obtained by comparing which mixture each latent variable point Z_j was in for each iteration. From this information an N by N matrix M was created displaying the probability of belonging with the other points. For example if we had two well separated mixtures and had 10 subjects each with a value in the latent variable, $Z_1, \ldots, Z_{10}$. Where $Z_1, \ldots, Z_5$ came from the first mixture and $Z_6, \ldots, Z_{10}$ came from the second mixture. We could obtain mixture membership vectors for each iteration, below is an example for 3 iterations.

Iteration 1 1,1,2,1,1,2,2,2,2,2

Iteration 2 1,1,1,1,1,2,2,2,2,2

Iteration 3 1,1,1,1,1,2,2,2,2,2

Here iterations 2 and 3 have discovered the correct mixture membership where as iteration 3 has one point that has been misclassified. Summarising the information in

these 3 iterations in a matrix M containing the probability of belonging in the same mixture as the other points, we obtain;

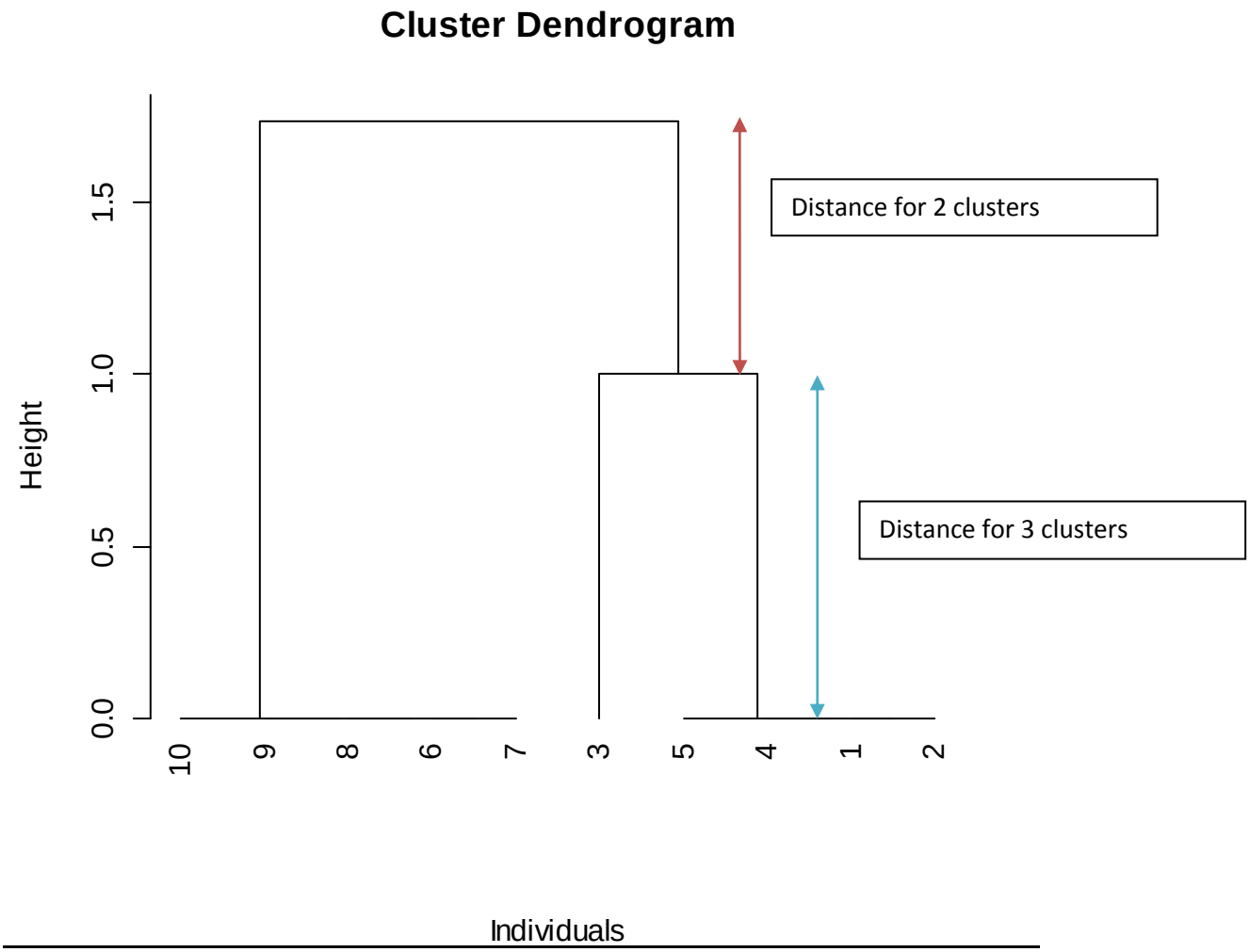
$$M = \begin{matrix} 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.67 & 0.67 & 1 & 0.67 & 0.67 & 0.33 & 0.33 & 0.33 & 0.33 & 0.33 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{matrix}$$

The matrix of probabilities M is then transformed into a matrix, P of probabilities of not belonging in the same mixture as all the other points Z_j , creating another N by N matrix. By computing P= 1-M we obtain;

$$P = \begin{matrix} 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0.33 & 033 & 0 & 0.33 & 0.33 & 0.67 & 0.67 & 0.67 & 0.67 & 0.67 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0.33 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0.67 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

This is then treated as a dissimilar measure between subjects, like Euclidean distance is used in classic clustering, but this new dissimilar metric is based on probabilities of belonging in the same group. The dissimilar metric is used to hierarchical cluster the subjects. The number of mixtures is determined by the number of mixtures with the largest distance in the dendrogram, see figure 6.6.3.1 obtained from hierarchical clustering of the probability of not belonging in a mixture with another subject.

Figure 6.6.3.1 Cluster memberships can be achieved by cutting the dendrogram at the level with the largest distance and achieving a strict partition of the data using the most probable number of clusters. For the example the largest distance between the clusters is found for two clusters; 1 cluster containing subjects 1,2,3,4 and 5 and cluster 2 containing subjects 6,7,8,9 and 10.



6.7 Simulations and Scenarios

For the six normally distributed variables simulations and scenarios were carried out for both 200 subjects and 500 subjects. For simulations with 2 binary variables and 4 normally distributed variables for 200 subjects were used to determine if the statistical model could return the correct number of clusters.

6.7.1For 6 normally distributed variables for 200 subjects

6.7.1.1 Scenario 1

The mixtures are far apart and are of equal size, ratio 1:1.

This indicates a good separation between the two groups which should be easy for the model to detect and partition well. The latent variable used to derive the inputs is simulated using two mixtures which are far apart and are in a 1:1 ratio.

$$Z \sim 0.5.\, N(0,2) + \ 0.5N(10,1) \qquad \textit{Equation 184}$$

Figure 6.7.1.1 .1The distributions of the latent variable Z used as an input to derive the Y values in each of the 10 different seeded simulations 1-10.

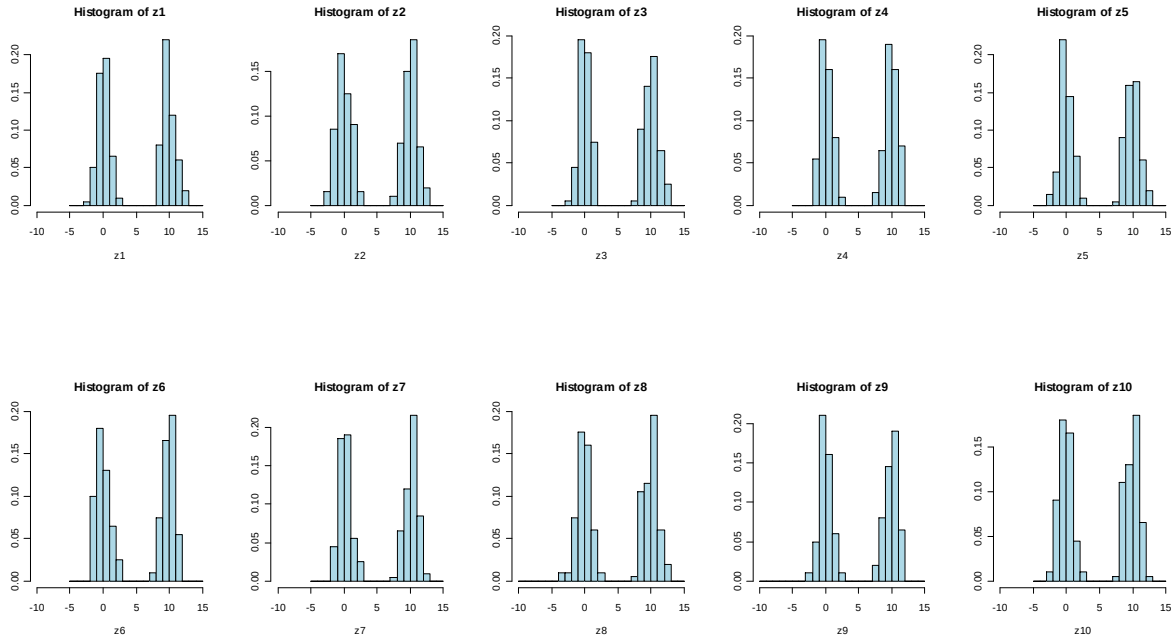


Figure 6.7.1.1.2. The graphs are the posterior distributions of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the above Graphs for the actual Latent variable used to derive the data it can be seen that the model has returned the underlying structure of the Latent Variable Z correctly.

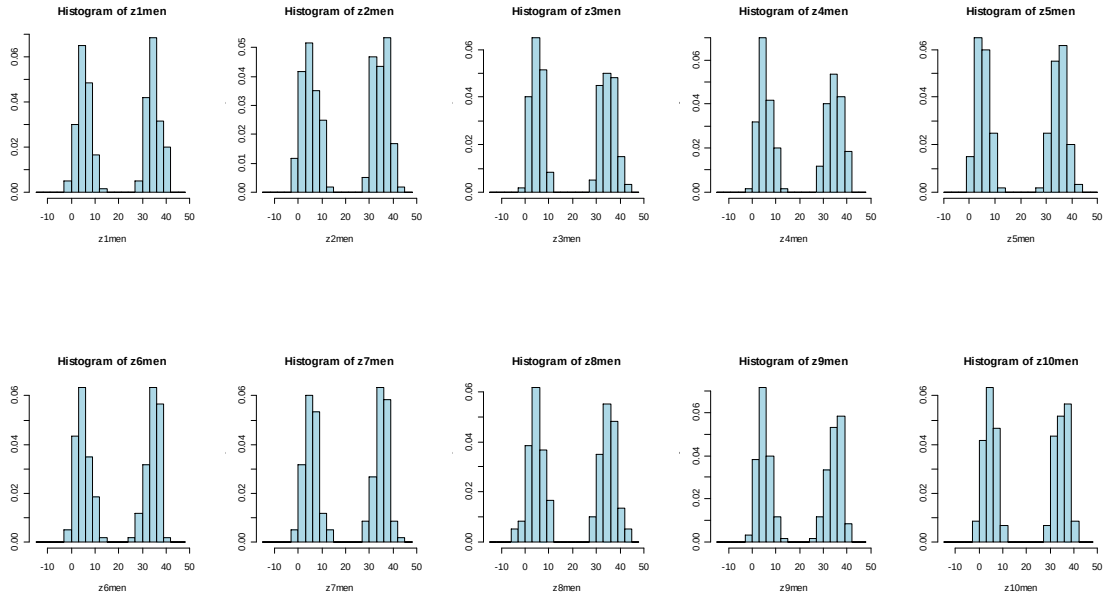


Table 6.7.1.1.1 below summarises the statistics for each iteration, as can be seen all iterations of each simulation pass the diptest at p=0.05 and each simulation has a large mean dip statistic indicating a multimodal distribution. The parameters converged well and the correct number of clusters was returned for all simulations with 100% correct cluster membership.

Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	100.00	2	100	100	0.1495
2	99.55	2	100	100	0.1369
3	100.00	2	100	100	0.1493
4	100.00	2	100	100	0.1448
5	100.00	2	100	100	0.1462
6	99.55	2	100	100	0.1440
7	100.00	2	100	100	0.1453
8	99.55	2	100	100	0.1416

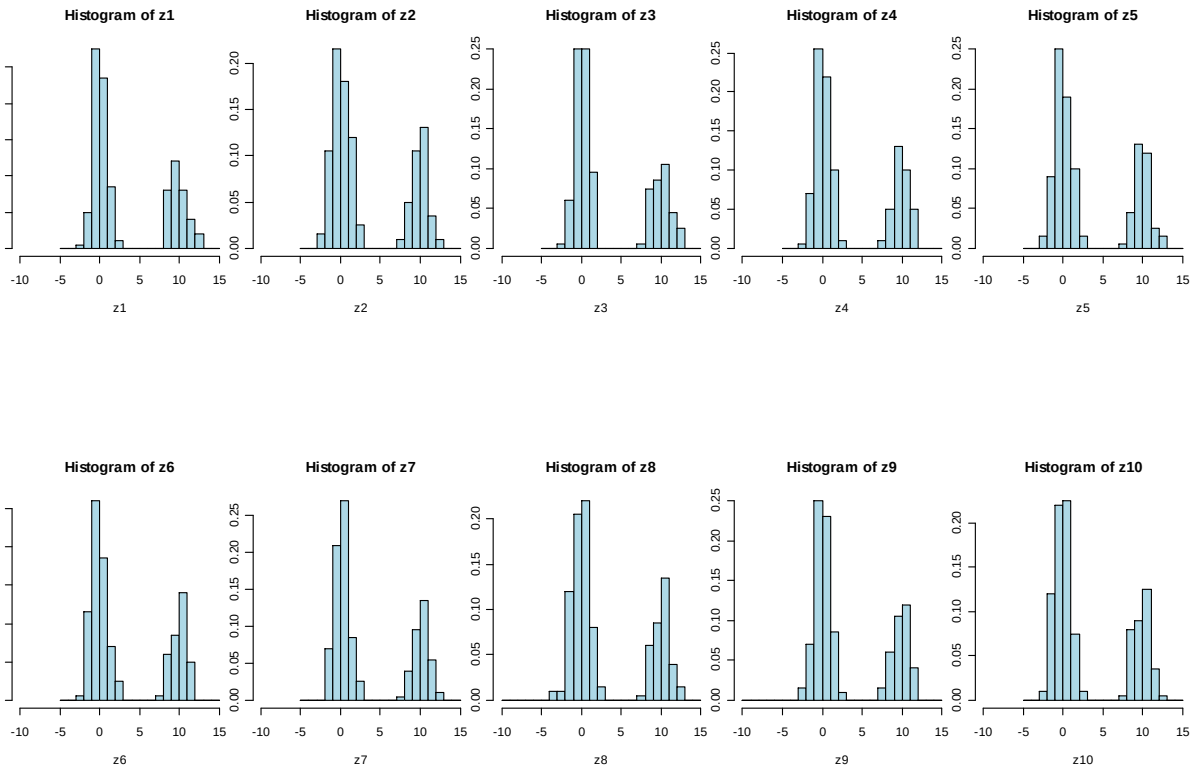
9	100.00	2	100	100	0.1457
10	99.55	2	100	100	0.1525

6.7.1.2 Scenario 2

This simulation describes a good separation between the two groups which should be easy for the model to detect and partition well. The latent variable used to derive the inputs is simulated using two mixtures which are far apart and are in a 2:1 ratio.

$$Z \sim 0.66.N(0,1) + 0.34N(10,1) \qquad \textit{Equation 185}$$

Graph 6.7.1.2.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



The graph 6.7.1.2.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly.

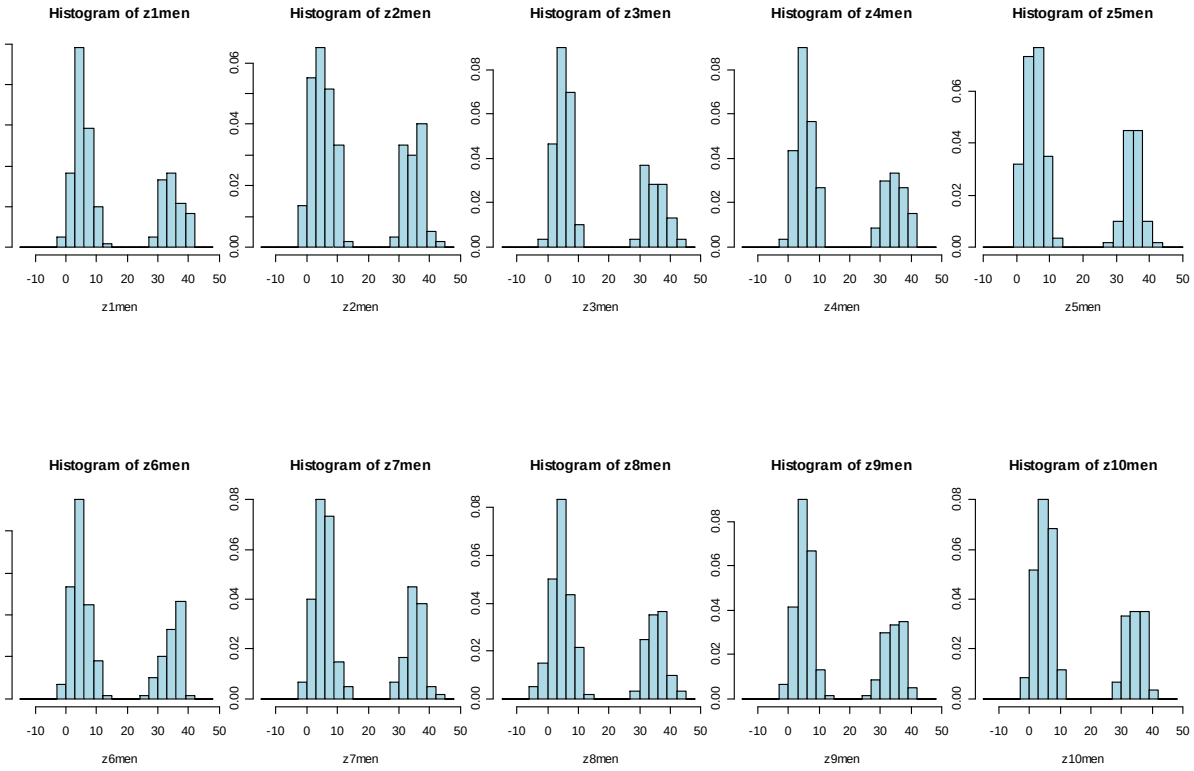


Table 6.7.1.2.1 below summarises the statistics for each iteration, as can be seen all iterations of each simulation pass the diptest at p=0.05 and each simulation has a smaller mean dip statistic then the previous scenario indicating that Z is a multimodal distribution but not to the extent of the previous simulation. The parameters

converged well and the correct number of clusters was returned for all simulations with 100% correct cluster membership.

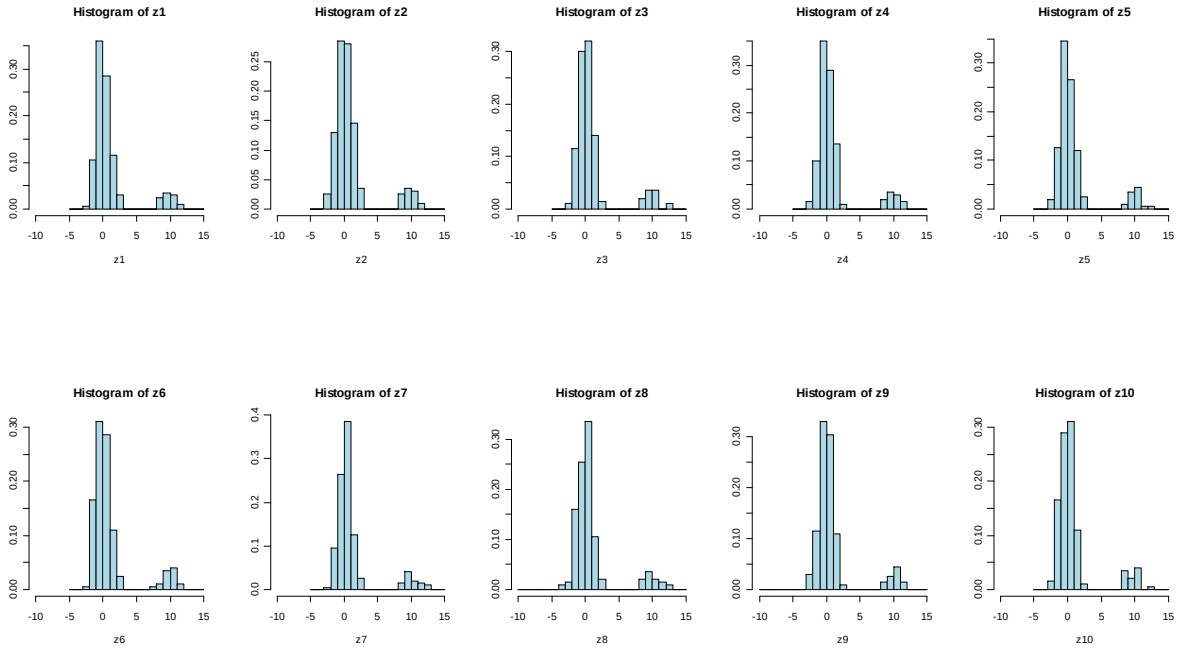
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	100.00	2	100	100	0.0991
2	99.09	2	100	100	0.1033
3	98.64	2	100	100	0.0984
4	100.00	2	100	100	0.0986
5	100.00	2	100	100	0.1052
6	100.00	2	100	100	0.1008
7	98.64	2	100	100	0.1012
8	99.55	2	100	100	0.0984
9	99.10	2	100	100	0.0977
10	100.00	2	100	100	0.1019

6.7.1.3 Scenario 3

This simulation describes a good separation between the two groups which should be easy for the model to detect and partition well. The latent variable used to derive the inputs is simulated using two mixtures which are far apart and are in a 9:1 ratio.

$$Z \sim 0.9.\, N(0,1) + \ 0.1N(10,1) \qquad \textit{Equation 186}$$

Graph 6.7.1.3.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



Graph 6.7.1.3.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly.

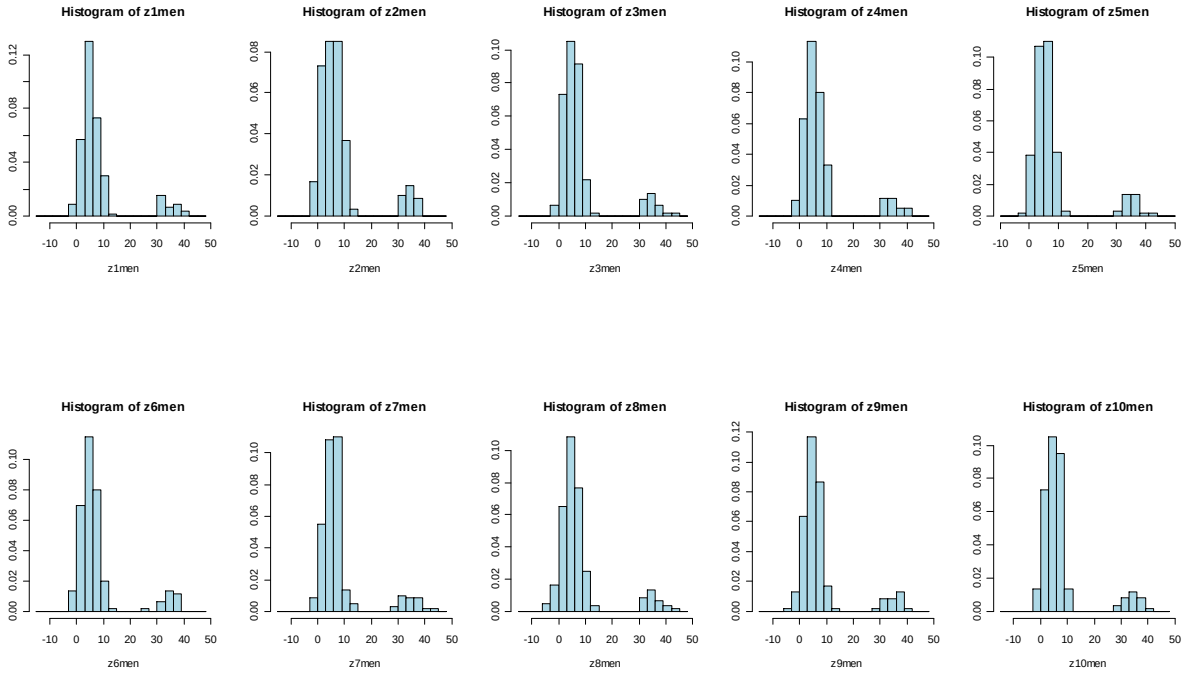


Table 6.7.1.3.1 below summarises the statistics for each iteration, as can be seen all iterations of each simulation pass the diptest at p=0.05 and each simulation has a smaller mean dip statistic then the last two scenario indicating a multimodal distribution but not to the degree of the first simulation. The parameters converged well and the correct number of clusters was returned for all simulations with 100% correct cluster membership.

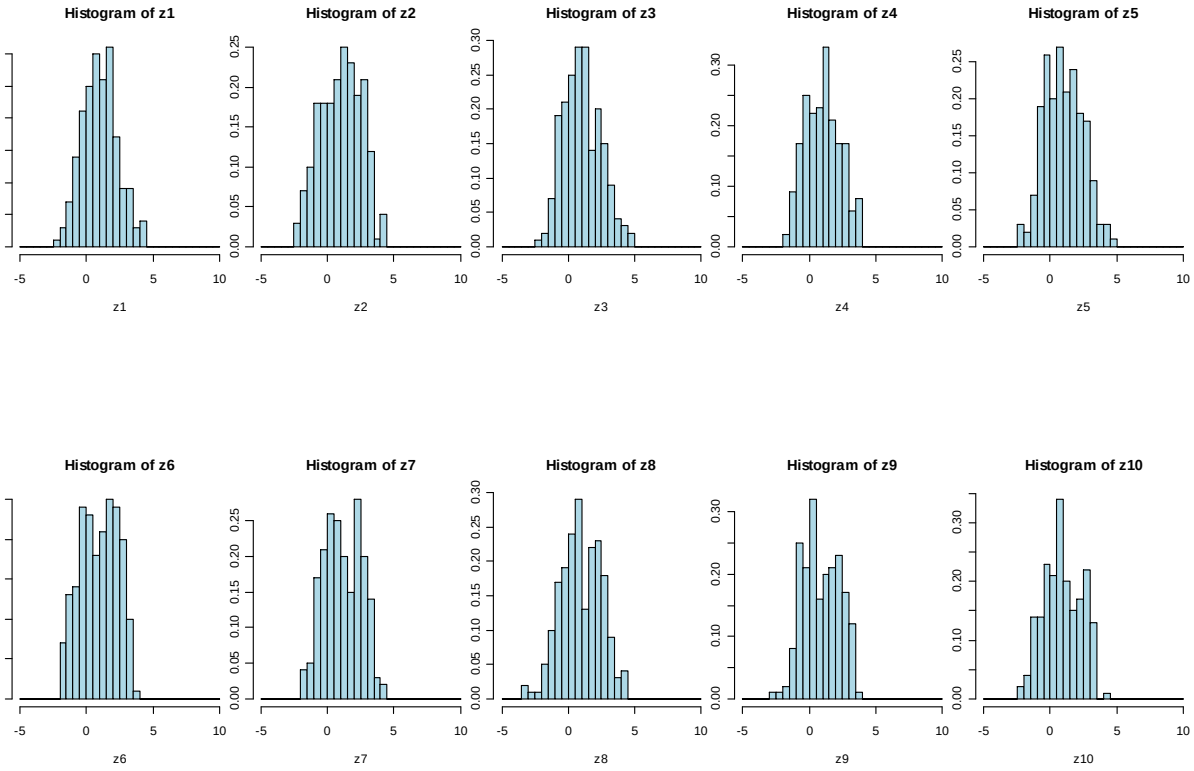
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	99.09	2	100	100	0.0322
2	100.00	2	100	100	0.0340
3	99.55	2	100	100	0.0318
4	99.09	2	100	100	0.0310
5	98.18	2	100	100	0.0330
6	100.00	2	100	100	0.0332
7	99.55	2	100	100	0.0271
8	99.09	2	100	100	0.0293
9	100.00	2	100	100	0.0300
10	99.55	2	100	100	0.0310

6.7.1.4 Scenario 4

This simulation describes a small separation between the two groups which could be difficult for the model to detect and partition well. The latent variable used to derive the inputs is created using two mixtures which are very close together and are in a 1:1 ratio.

$$Z \sim 0.5.\, N(0,1) + \ 0.5N(2,1) \qquad \textit{Equation 187}$$

Graph 6.7.1.4.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10.



Graph 6.7.1.4.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent (Dorazio 2009)see clear sub-groups as the mixtures overlap.

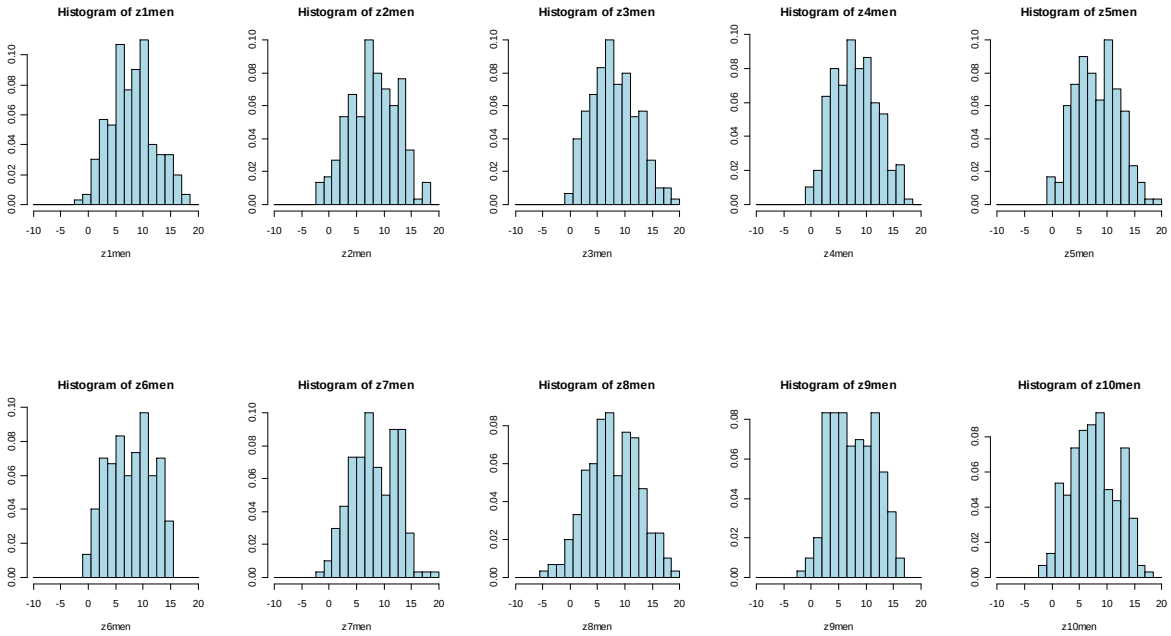


Table 6.7.1.4.1 below summarises the statistics for each iteration, some of the iterations for simulations pass the diptest but others do not indicating that there is uncertainty whether the distribution is multimodal or not. Each simulation has a smaller mean dip statistic then the last three scenarios indicating less faith in the hypothesis of a multimodal distribution. However the parameters converged well and the correct number of clusters was returned for all simulations and cluster membership was greater than 56.50 % of the iterations for all simulations.

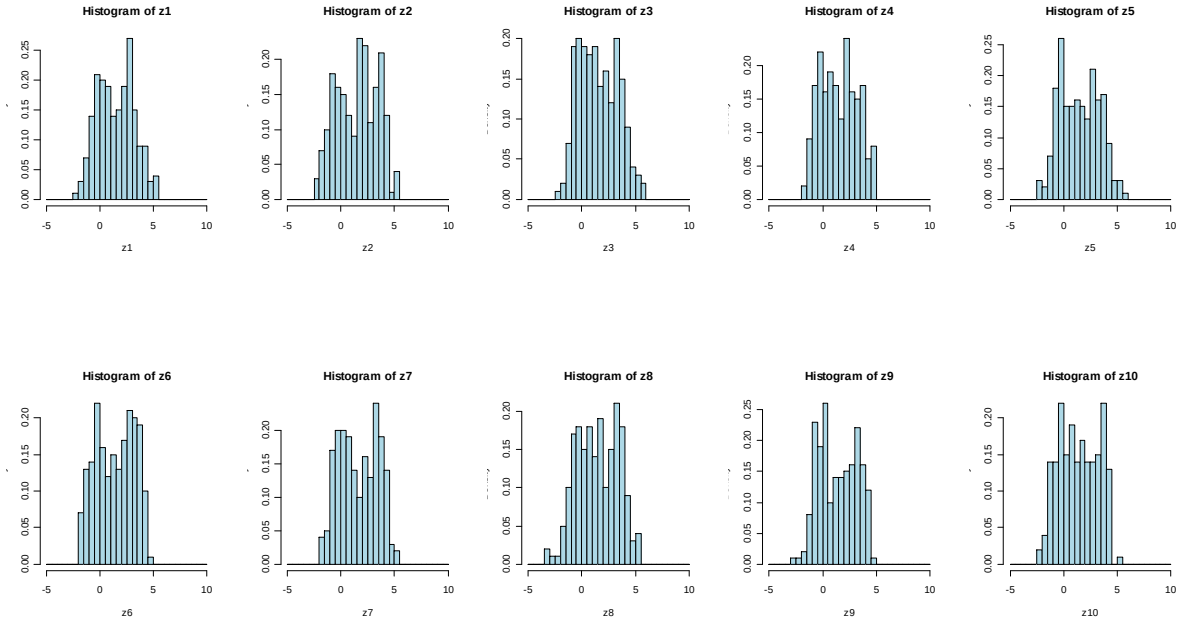
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	99.55	2	56.50	65.05	0.0204
2	99.10	2	83.50	66.62	0.0203
3	99.55	2	69.50	49.96	0.0188
4	99.09	2	78.00	67.25	0.0203
5	99.09	2	83.50	82.74	0.0224
6	99.55	2	80.50	85.12	0.0226
7	98.18	2	81.50	98.04	0.0278
8	98.18	2	85.50	70.36	0.0206
9	100.00	2	79.50	92.35	0.0241
10	99.55	2	74.00	70.91	0.0206

6.7.1.5. Scenario 5

This simulation describes a slightly larger separation between the two groups then the last simulation which should be easier for the model to detect and partition then the last scenario. The latent variable used to derive the inputs is created using two mixtures which are fairly close together and are in a 1:1 ratio.

$$Z \sim 0.5.\, N(0,1) + \ 0.5N(3,1) \qquad \textit{Equation 188}$$

Graph 6.7.1.5.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



Graph 6.7.1.5.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has

returned the underlying structure of the Latent Variable Z correctly, but it is still slightly difficult to see clear sub-groups as again the mixtures overlap in some of the simulations.

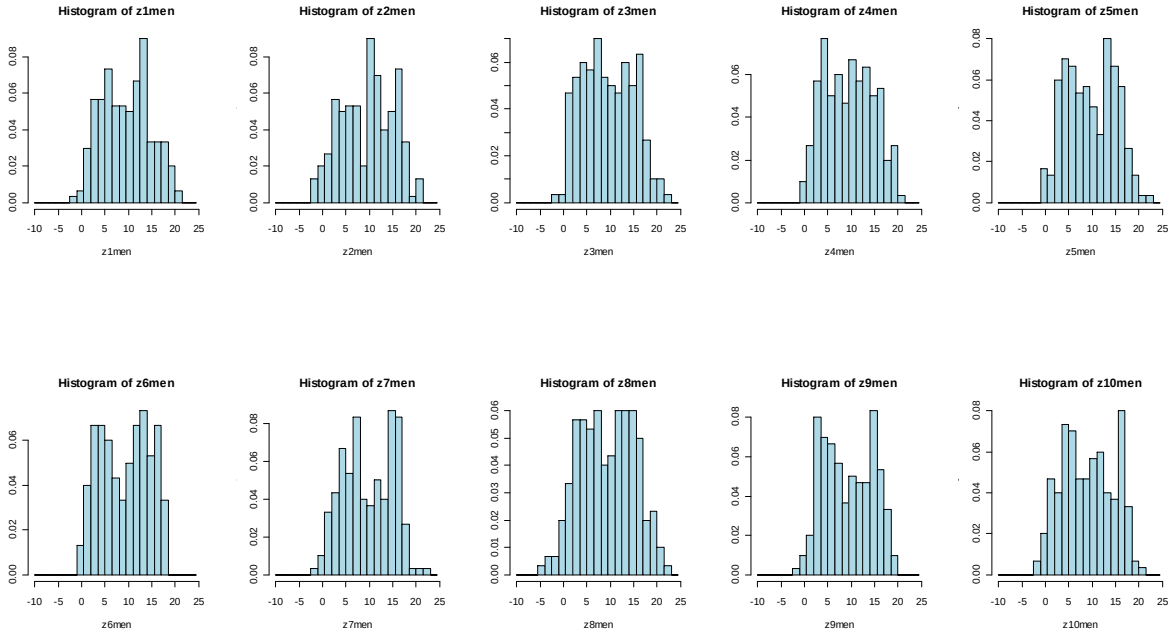


Table 6.7.1.5.1 below summarises the statistics for each iteration, all iterations of each simulation have a high percentage of passing the diptest at p=0.05 and each simulation had a larger mean dip statistic then the last scenario which had the groups mean closer together. This indicated more faith in a multimodal distribution then the last scenario. The parameters converged well and the correct number of clusters was returned for all simulations with high correct cluster membership greater than 88% for all simulations.

Simulation	Percentage	Number of	Percentage	Percentage of	Mean of

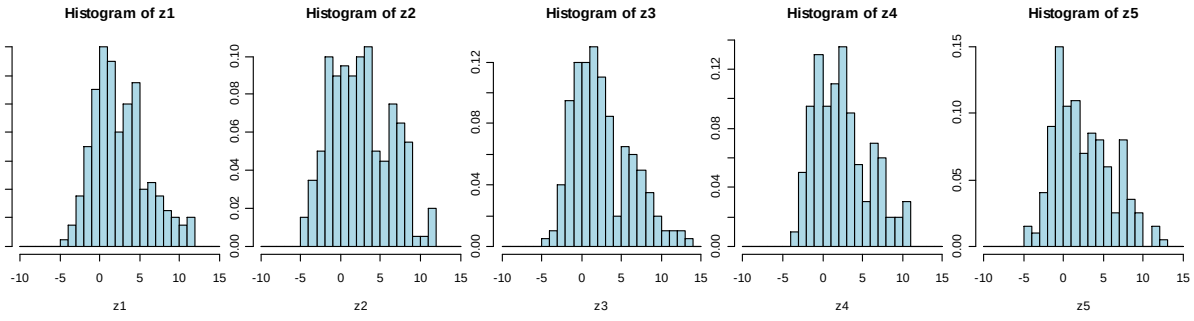
Number	of parameters passing Heidelberg and Welch diagnostic	mixtures found for maximum distance in dendrogram	correct cluster membership	iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	dip statistic for all iterations
1	99.55	2	92.50	96.66	0.0262
2	99.09	2	91.00	97.58	0.0256
3	99.09	2	93.50	85.63	0.0221
4	100.00	2	91.50	94.04	0.0243
5	99.55	2	88.00	99.94	0.0327
6	99.09	2	89.50	99.97	0.0328
7	97.73	2	90.50	99.99	0.0357
8	99.55	2	90.00	97.34	0.0246
9	99.55	2	92.00	100.00	0.0356
10	99.09	2	95.50	92.29	0.0233

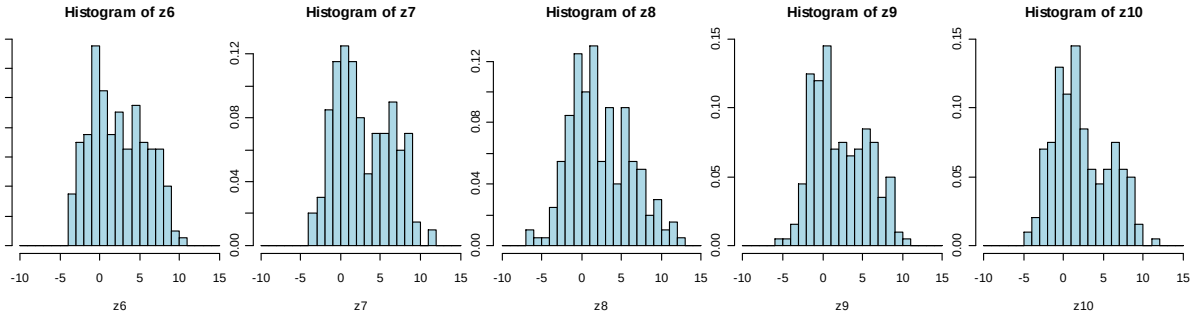
6.7.1.6 Scenario 6

This scenario describes a slightly larger separation between the two groups then the last scenario; it also has a larger standard deviation for one of the groups which could affect correct cluster membership as the two groups overlap. The latent variable used to derive the inputs is created using two mixtures which are fairly close together and are in a 1:1 ratio.

$$Z \sim 0.5.N(0,2) + 0.5N(5,3) \qquad \textit{Equation 189}$$

Graph 6.7.1.6.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10





Graph 6.7.1.6.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly, but it is difficult to see clear sub-groups as again the mixtures overlap.

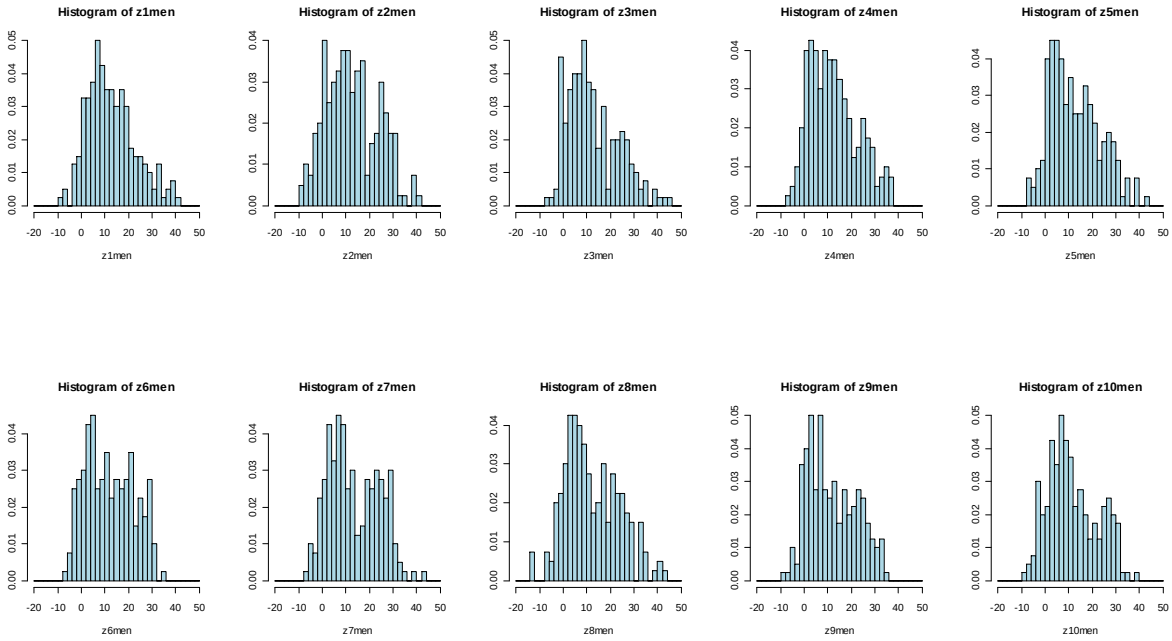


Table6.7.1.6.1 below summarises the statistics for each iteration. The iterations of each simulation have a mixed percentage of passing the diptest at p=0.05 form 35%-99.9% depending on the simulation seed used, each simulation had a smaller mean dip statistic then all the other previous scenarios. This indicated less faith in a multimodal distribution then the previous simulations. The parameters converged well and the correct number of clusters was returned for all simulations with high correct cluster membership, greater than 74% for all simulations.

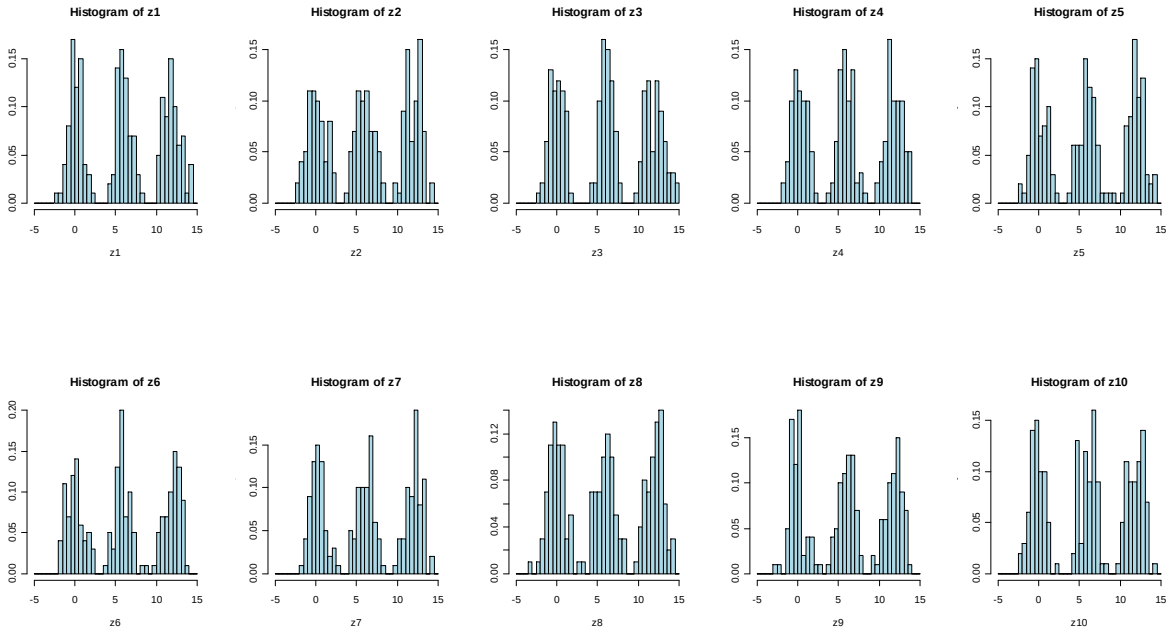
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	99.55	2	84.00	35.75	0.0178
2	99.55	2	74.50	89.44	0.0216
3	99.55	2	76.00	67.54	0.0198
4	100.00	2	81.00	59.60	0.0193
5	99.09	2	85.00	34.01	0.0177
6	100.00	2	81.00	78.24	0.0209
7	100.00	2	82.00	99.99	0.0286
8	100.00	2	82.50	46.24	0.0184
9	100.00	2	83.00	69.62	0.0200
10	99.54	2	77.00	90.02	0.0216

6.7.1.7 Scenario 7

This scenario describes a large separation between three groups. The latent variable used to derive the inputs is created using three mixtures which are fairly close together and are in a 1:1:1 ratio.

$$Z \sim 0.33.N(0,1) + 0.33N(6,1) + 0.34.N(6,1) \qquad \textit{Equation 190}$$

Graph 6.7.1.7.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



Graph 6.7.1.7.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly and the three sub-groups are easily visible.

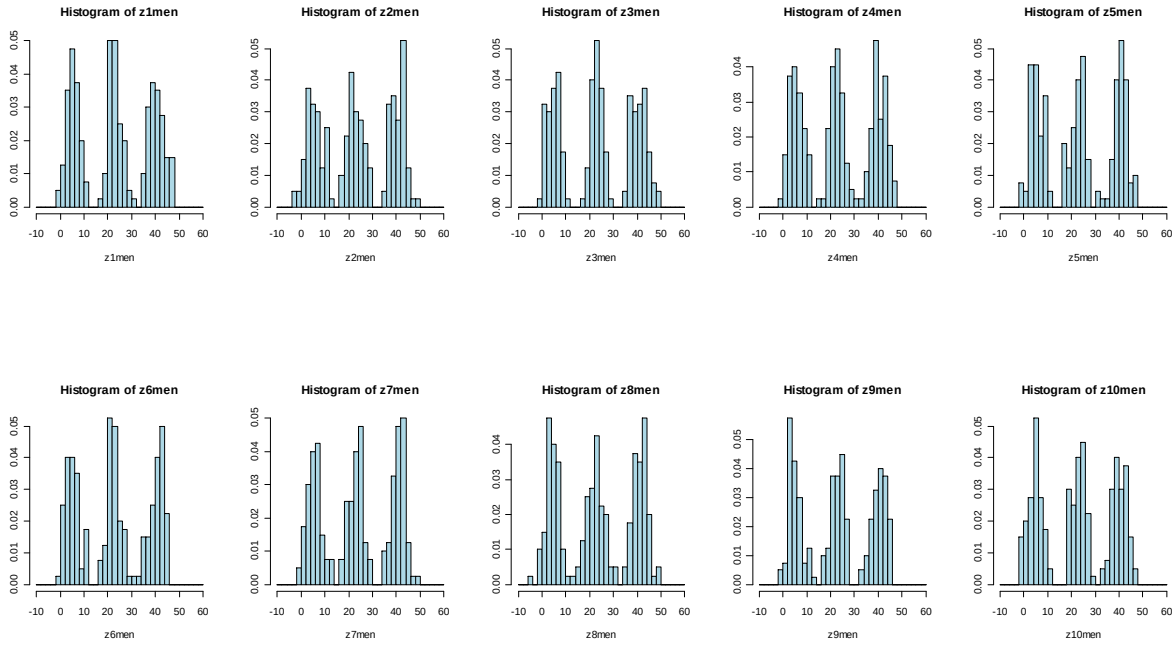


Table 6.7.1.7.1 below summarises the statistics for each simulation. The iterations of each simulation pass the diptest at p=0.05, 100% of the time. Each simulation has a relatively large mean dip statistic then the previous scenarios which have less obvious clusters which overlap. A large dip statistic indicated greater faith in a multimodal distribution being a correct solution. The parameters converged well and the correct

number of clusters was returned for all simulations with very high correct cluster membership of greater than 98% for all simulations.

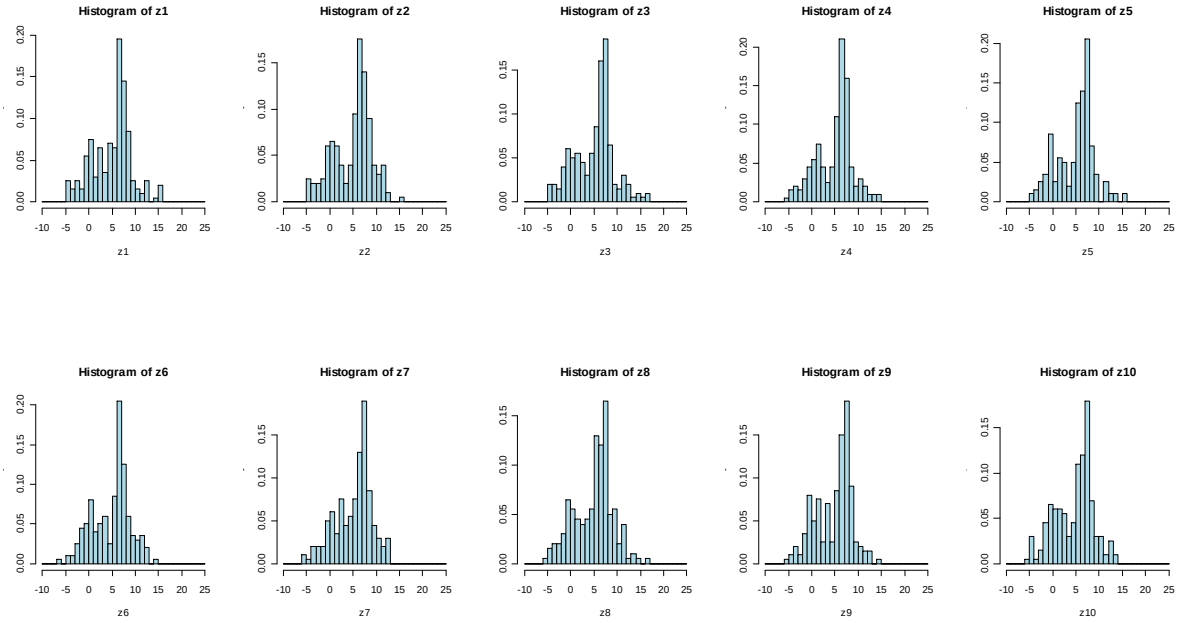
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	100.00	3	99.50	100	0.0739
2	98.64	3	100.00	100	0.0620
3	99.09	3	100.00	100	0.0771
4	99.55	3	99.50	100	0.0676
5	100.00	3	99.00	100	0.0773
6	100.00	3	99.50	100	0.0706
7	99.09	3	100.00	100	0.0718
8	99.55	3	98.50	100	0.0675
9	100.00	3	100.00	100	0.0730
10	100.00	3	100.00	100	0.0714

6.7.1.8. Scenario 8

This scenario describes small separations between five groups. The latent variable used to derive the inputs is created using five mixtures which are fairly close together and are in a 1:2:4:2:1 ratio.

$$Z \sim 0.1.N(0,1) + 0.2.N(3,2) + 0.4N(7,1) + 0.2N(9,3) + 0.1N(-2,2) \qquad \textit{Equation 191}$$

Graph 6.7.1.8.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



Graph 6.7.1.8.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly although the five mixtures are difficult to see in both the original Z and the posterior Z. Two groups are roughly visible however.

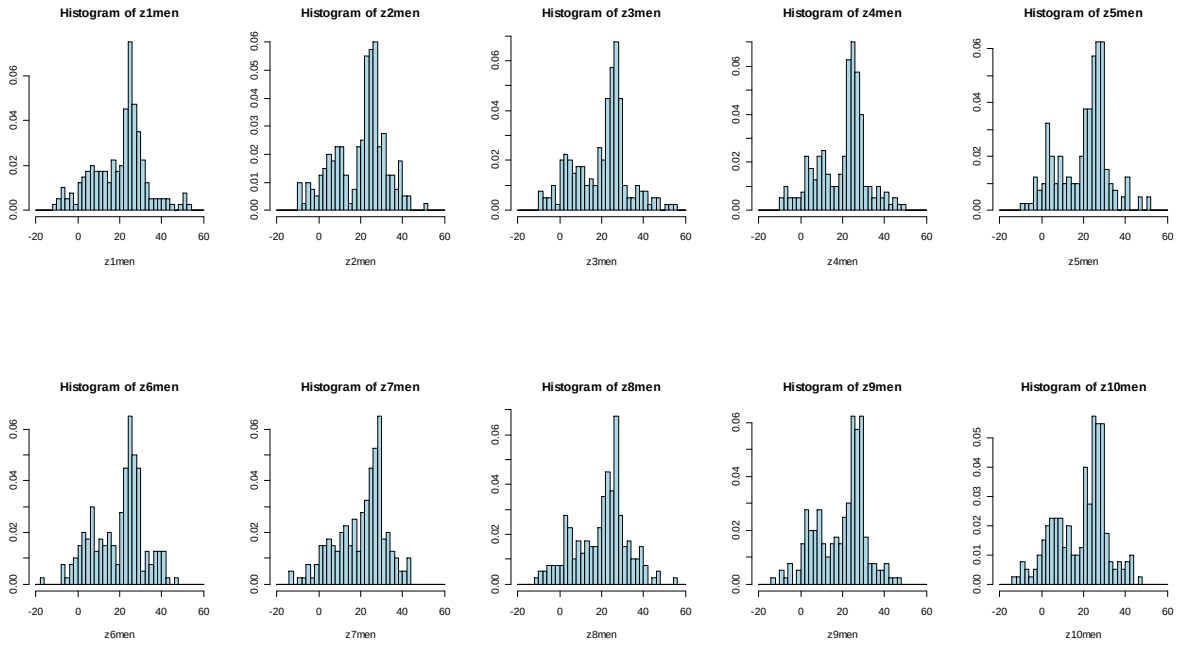


Table 6.7.1.8.1 below summarises the statistics for each simulation. Most simulations pass the diptest at p=0.05, with a high percentage of iterations passing each simulation however two of these simulations have a very low percentage pass rate. This is demonstrated by the two simulations having smaller mean dip statistics and their posterior distributions looking unimodal. The parameters converged well but the

correct number of clusters was not returned however. Two clusters were returned that indicate the two modes visible in some of the posterior graphs.

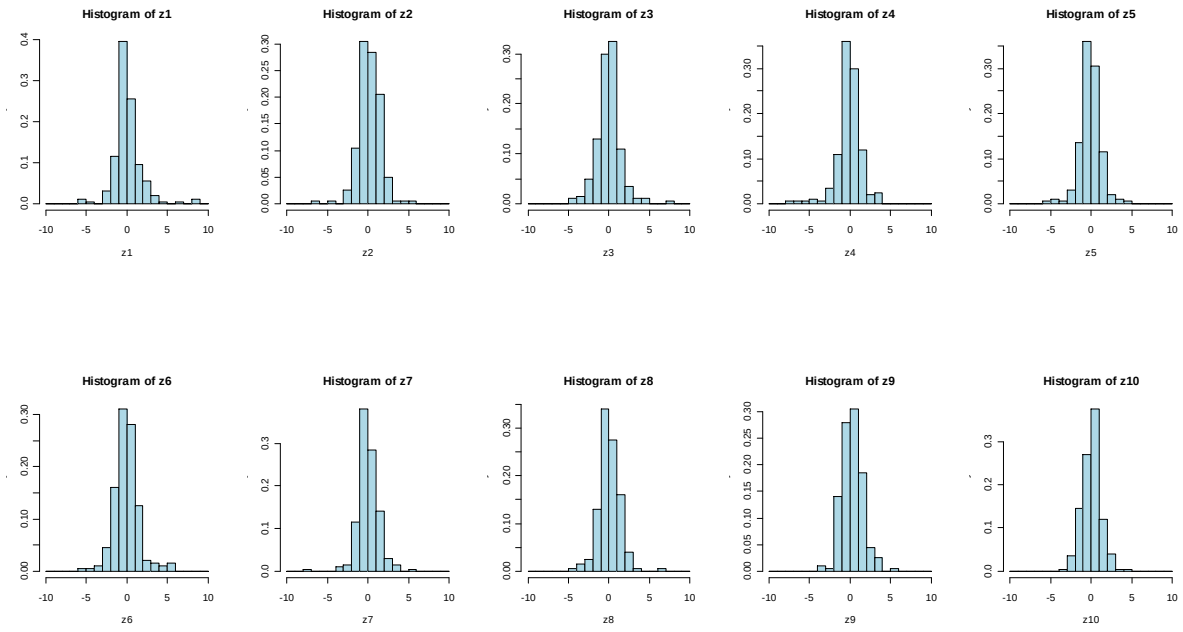
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	99.55	2	NA	22.06	0.0169
2	99.090	2	NA	99.99	0.0285
3	100.00	2	NA	99.13	0.0238
4	100.00	2	NA	99.92	0.0263
5	98.18	2	NA	99.98	0.0276
6	100.00	2	NA	94.61	0.0223
7	100.00	2	NA	19.69	0.0167
8	99.09	2	NA	74.88	0.0202
9	99.09	2	NA	99.98	0.0266
10	99.55	2	NA	99.98	0.0265

6.7.1.9 Scenario 9

This scenario describes a non-normal distribution with no mixtures. The latent variable used to derive the inputs is created using a t distribution with 4 degrees of freedom.

$$Z \sim t_{dist}(4) \qquad \textit{Equation 192}$$

Graph 6.7.1.9.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



Graph 6.7.1.9.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly.

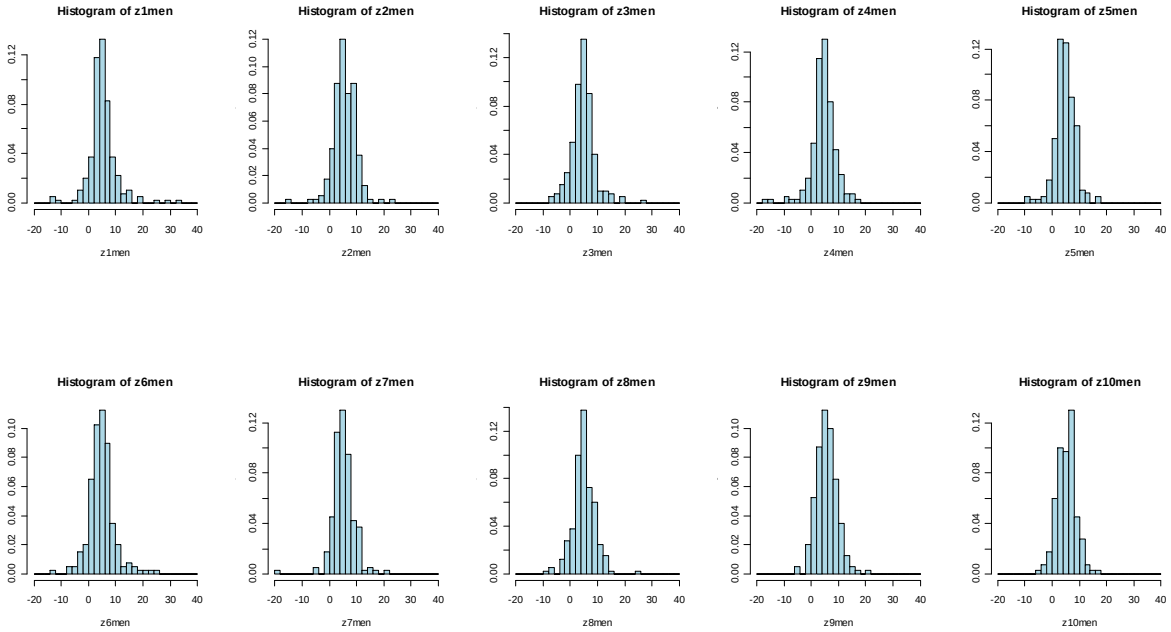


Table 6.7.1.9.1 below summarises the statistics for each simulation. Most simulations do not pass the diptest at p=0.05, with percentage of iterations passing between 29-62%. This is demonstrated by the simulations having a very small mean dip statistics and their posterior distributions looking unimodal. The parameters converged well but the correct number of clusters was not returned however as the cluster algorithm used cannot return a value for 1.

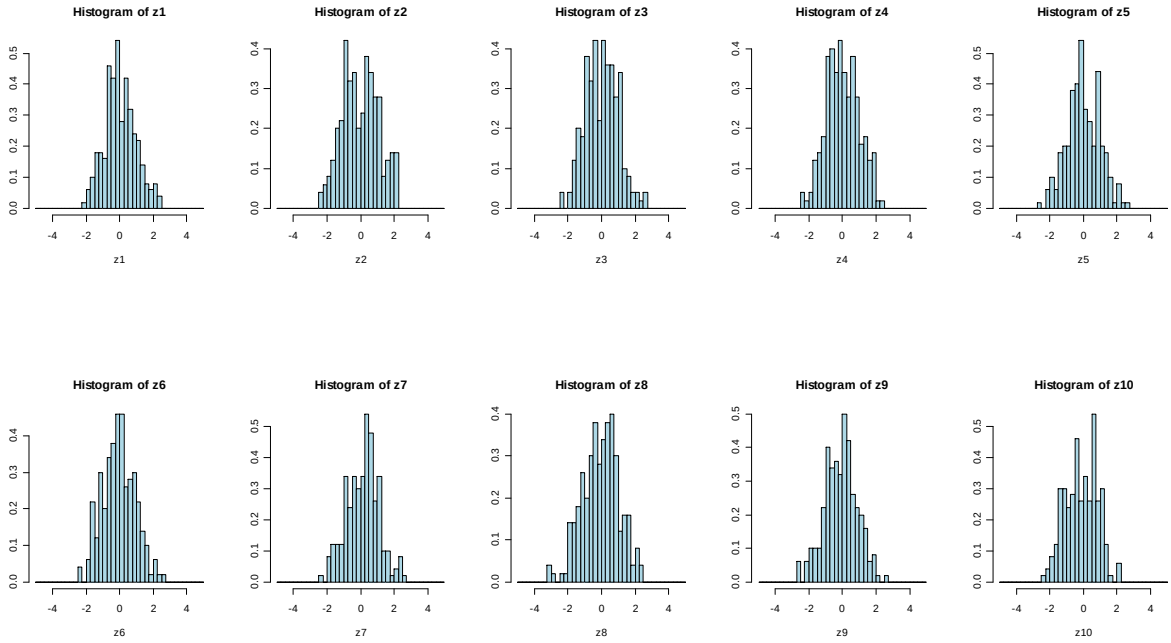
Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	100.00	2	NA	29.56	0.0172
2	98.64	2	NA	55.49	0.0194
3	100.00	2	NA	40.98	0.0182
4	100.00	2	NA	44.96	0.0185
5	100.00	2	NA	47.55	0.0188
6	100.00	2	NA	61.40	0.0200
7	100.00	2	NA	34.85	0.0177
8	100.00	2	NA	35.80	0.0178
9	99.09	2	NA	42.28	0.0183
10	99.55	2	NA	61.75	0.0201

6.7.1.10 Scenario 10

This scenario describes a normal distribution with no mixtures. The latent variable used to derive the inputs is created using a standard normal distribution as in Classical latent variable models.

$$Z \sim N\ (0, 1) \qquad \textit{Equation 193}$$

Graph 6.7.1.10.1 below demonstrates the distribution of the latent variable Z used as an input to derive the Y values in each of the 10 simulations 1-10



Graph 6.7.1.10.2 below demonstrates the posterior distribution of the mean (zmen) of the values of the latent variable Z found using the Dirichlet Process Mixture Latent Variable Model for each of the 10 simulations 1-10. By comparing with the Graphs for the actual Latent variable used to derive the data above we can see that the model has returned the underlying structure of the Latent Variable Z correctly.

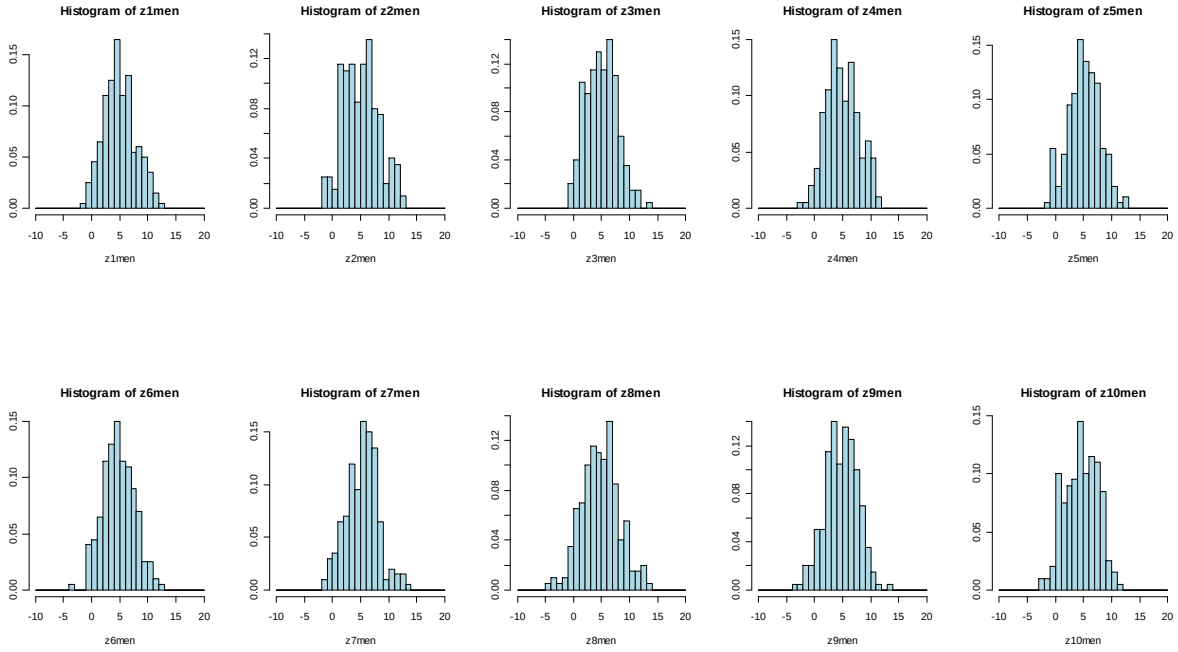


Table 6.7.1.10.1 below summarises the statistics for each simulation. Most simulations do not pass the diptest at p=0.05, with percentage of iterations passing between 49-80%. This is demonstrated by the two simulations having very small mean dip statistics and their posterior distributions looking unimodal. The parameters converged well but the correct number of clusters was not returned however as the cluster algorithm used cannot return a value for 1.

Simulation Number	Percentage of parameters passing Heidelberg and Welch diagnostic	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	99.09	2	NA	52.63	0.0192
2	100.00	2	NA	80.03	0.0224
3	100.00	2	NA	74.60	0.0214
4	99.09	2	NA	64.34	0.0204
5	99.55	2	NA	49.73	0.0190
6	98.64	2	NA	58.12	0.0197
7	100.00	2	NA	52.83	0.0192
8	100.00	2	NA	59.91	0.0198
9	99.54	2	NA	68.43	0.0208
10	60.45	2	NA	75.44	0.0212

Table 6.7.1.10.2 indicates the mean dip statistic needed to establish unimodality/ multimodality for data with 200 individuals over 10 scenarios.

Mean dip statistic for simulations (N=200)	Annotation	Corresponding p-levels on table
≤0.019	Uni-modal	≤0.05
0.019 to 0.023	Uni-Modal or Overlapping Clusters with Low Percentage of Correct Cluster Membership	0.05 to 0.3
≥0.023	Clusters detected with High Percentage of Correct Cluster Membership	≥0.3

6.7.2. For 6 normally distributed variables and 500 subjects

The 10 scenarios were then simulated ten times using different seeds again but this time for 500 subjects, to determine if the statistical modelling was consistent when increasing the number of subjects, the results for these scenarios can be found in appendix4 .

Mean dip statistic for simulations (N=500)	Annotation	Corresponding p-levels on table
0.012	Uni-modal	≤0.05
0.012 to 0.013	Uni-modal or Overlapping Clusters with Low Percentage of Correct Cluster Membership	0.05 to 0.1
≥0.012	Clusters detected with High Percentage of Correct Cluster Membership	≥0.1

6.7.3. For 4 Normal variables and 2 Binary variables

The data simulated describe the same scenarios for the 6 normally distributed variables but in this case the latent variable scenarios are used to simulate 4 normal variables and 2 binary variables using 200 subjects. The results of these simulations can be found in appendix 5

6.8 Discussion

For most scenarios that originally contained mixtures the model returned the correct number of mixtures accompanied with a high percentage of correct cluster membership for the clusters. The percentage of correct cluster membership did not seem to be effected by the size or proportions of the mixtures. However distance between the means of the mixtures and the size of the standard deviation of the mixtures did affect the percentage of correct cluster membership and mixture detection. As expected mixtures that were far apart had a higher percentage of

correct cluster membership then clusters close together, this is due to overlapping of the mixtures as for some points in mixtures that overlap it is impossible to say which mixture they belong to.

Wrong numbers of clusters can be returned when two mixtures overlap and especially when the mixtures overlap such that they appear to look like 1 distribution when plotted. This is found in scenario 8 where we have 5 mixtures and only two were detected. Although this looks like a failure of the model on inspection it was found that original mixtures 1, 2 and 5 had been merged into 1 cluster and mixtures 3 and 4 were merged into the other cluster as mixtures 1, 2 and 5 were too close together, as were 3 and 4 and it was impossible to separate them just by their original distribution so the distribution could equally be simulated by the two clusters found. This is reflected in the simulations low mean dip statistic and the plots of both original distribution and posterior.

It is worth noting that most of the mixtures that were correctly found related to a higher mean dip statistic rejecting the null hypothesis of a unimodal distribution. For scenarios 9 and 10 where the latent variable distribution is unimodal the model wrongly described them as containing 2 mixtures, this is due to the nature of the hierarchical clustering algorithm that determines the best number of clusters not equal to 1. For these two scenarios the mean dip statistic was always below 0.023 (for 200 subjects) and 0.012 (for 500 subjects).

For scenario 8 containing the mixtures that overlapped with each other the dip statistic for each simulation was also small with all simulations having dip statistic below 0.029 (for 200 subjects) and 0.024 (for 500 subjects). For scenarios 1-5 the mean dip statistic

is always above 0.018 (for 200 subjects) and 0.011 (for 500 subjects). If a cut off around the 0.023 (for 200 subjects) or 0.013 (for 500 subjects) mark is used we can begin to accept the cluster number and the cluster membership derived from that distribution, if it is below 0.018 (for 200 subjects) and below 0.012 (for 500 subjects) however we can differently say the distribution is uni-modal.

Binary data variables mixed with normal distributed variables performed well and returned similar results to the normally distributed scenarios in all cases. However it is unclear whether the latent variable structure was heavily dependent on the 4 remaining normal variables rather than the 2 binary ones. Meaning that the binary variables might add little information as it is impossible to determine the underling structure from such binary data alone. Also the coefficients that were associated with the binary variables often failed the Heidelberg and Welch diagnostic and on inspection seemed bimodal suggesting that these were not converged or that they were non-normal.

6.9 Conclusion

The model can accurately determine the latent variable distribution shape that is created from mixtures of normal distributions (multi-modal) or created from non-normal/ normal distributions (uni-modal). Latent variable distribution shapes are detected using all normal variables and a mixture of normal and binary variables. Although using binary data may not be detrimental in the model as the latent variables are derived from the other normally distributed variables only.

If the latent variable contains mixtures the model can obtain the number of mixtures in the latent variable distribution for distributions with a mean dip statistic above 0.023

for 200 subjects or 0.012 for 500 subject's simulations with high dip statistics also have a higher correct cluster membership. If the mean dip statistic for a latent variable distribution is below 0.023 then the number of clusters could be an approximation and the distribution is more likely to be of a uni-modal shape which could be made up of clusters or not, if the dip statistic is below 0.018 it is generally a uni-modal distribution.

These values can be used to determine the significant values to use on the table for other data with different amounts of subjects greater than 200. If the data have a mean dip statistic less than the table value at p=0.30, then it is unclear whether the data is uni-modal or not we can conclude that the cluster membership should be ignored. However if the data have a mean dip statistic greater than the one at p=0.30 then clusters are likely to be present and the cluster membership is likely to be high. It is worth noting that this represents an upper limit and the actual limit might be smaller especially when data has a large amount of subjects. If we take the traditional frequentist cut off at p=0.05 then we can definitely say if a distribution is uni-modal or not.

For the severe asthma data it is important to use normally distributed variables where possible as the binary variables would add little or no information to the underlying distribution. If clusters are detected in the severe asthma data their mean dip statistic should also be checked to act as a measure of how much we believe in the number of clusters obtained and the cluster membership for subjects.

6.10 Closing Chapter

Scenarios of different cluster types were simulated and tested to determine if the Dirichlet process normal latent variable model could determine these clusters and return correct cluster membership. The model proved useful in determining the number of clusters and cluster membership for data that contained clusters and was multi-modal. The model also highlighted the need to check distributions for multimodality as data that only contained one group i.e. no clusters, returned two clusters indicating an approximation of the distribution.

Chapter 7. Simulation for Correlated and Uncorrelated Outcomes

7.1 Chapter overview

Here simulations were carried out to determine if the Dirichlet process normal mixture latent variable model, DPNMLVM could be carried out over multiple independent latent variables using a number of measured variables that had differing correlations with each other. 5 scenarios were tested with differing mixtures over the differing number of latent variables. Solutions were found that were comparable to the original solutions indicating that the mixtures applied to one latent variable in chapter 6 could be applied for multiple latent variables.

7.2 Introduction

Having tested the Dirichlet process using correlated variables we now test that the model works when the dataset contains subsets of variables that are correlated within the subset but are not correlated with another subset, i.e. independent groups of variables. In this case we have several variables all being correlated to a number of latent variables or factors. The model needs to determine the clustering on each factor in a similar way to using correlated outcomes and we also need a method to determine the best number of factors to use in our model. There are infinitely many possible scenarios to test involving different factors and different mixtures on factors to show a range of these 9 continuous variables were simulated for 5 different scenarios. Each scenario uses a differing number of factors that are needed to be determined along with differing mixtures that are on each independent factor.

The scenarios are carried out to determine if the Dirichlet Process Normal Mixture Latent Variable Model, DPNMLVM can detect the variation seen in the scenarios. Once the models were ran and convergence obtained each factor was treated independently and the number of mixtures obtained for each factor was determined along with the dip statistic for each factor distribution, In a similar way to the simulations carried out previously for only one latent variable in chapter 6.

Statistics were taken from the iterations to determine

- If the mixtures could imply a multi modal distribution for each factor,

  This was derived from a frequentist hypothesis test for multi modality of static data called the dip test; a mean dip test for the latent variable was obtained by taking the mean of all the iterations.
- Whether the correct number of mixtures could be obtained from the model for each factor,

  Using hierarchical clusters of the probability of not belonging in a group with any other subject.
- Whether the correct cluster membership could be obtained for each factor,

  This was defined by cutting the dendrogram obtained from the probability clustering above at the point of maximum difference between clusters.

7.3 Generating simulations

For latent variables or factors that aren't correlated i.e. those found in scenarios 2, 4, 5 variables are generated using the following mixture methodology to create a latent variable distribution.

$$Z_{li} \sim \pi_{lk}.N(\mu_{lk}\ , {\sigma_{lk}}^2) \qquad \textit{Equation 194}$$

$$\sum_{k=1}^{M} \pi_{lk} = 1 \quad for\ l = 1,2, \ldots..L \qquad \textit{Equation 195}$$

Where M is the number of mixtures and π_{lk} is the proportion of the *k* normal distribution in the l latent variable Z_l, and $N(\mu_{lk}\ , \sigma_{lk})$ is a normal distribution with mean μ_{lk} and standard deviationσ_{lk}.

The nine continuous variables are generated using the l latent variables above in the formula below.

$$Y_{ij} \sim N\left(\mu_j, {\sigma_j}^2\right) \qquad \textit{Equation 196}$$

$$\mu_j = \beta_{0j} + \beta_{lj}.Z_{li} \qquad \textit{Equation 197}$$

Where $\alpha_j, \beta_j, \sigma_j$ are selected integers held constant over the simulations and scenarios, $Y_{ij}$is a measured variable for subject i and variable j distributed by$\left(\mu_j\ , \sigma_j\right)$ a normal distribution with mean μ_j and standard deviationσ_j . Values of $\beta_{0j}\ , \beta_{lj}, \sigma_j$ were chosen to represent a mixed variety of possible variables on differing scales and magnitudes. These are specified explicitly for each of the 9 variables for each of the scenarios; see tables 7.7.1, 7.8.1, 7.9.1, 7.10.1, 7.11.1 for scenarios 1, 2, 3, 4, 5 respectively.

For correlated variables the variables were derived in a similar way but this time the latent variables were derived from multivariate normal's using covariances.

7.4 Latent variable model used

For Normal Variables and subjects i for i=1, 2...N

$$Y_{ij} \sim N\left(\mu_{ij}, {\sigma_j}^2\right) \qquad for\ j = 1,2 \dots M \qquad \textit{Equation 198}$$

$$\mu_{ij} = \beta_{0j} + \beta_{lj}.Z_{li} \ \ for\ l = 1,2 \dots L \qquad \textit{Equation 199}$$

$$Z_{li} \sim D_l(\alpha_l, G_{lo}) \qquad \textit{Equation 200}$$

$$G_{lo} \sim N(\theta_{li}, V_{li}) \qquad \textit{Equation 201}$$

Where Y_{ij} represents the i individual of the j normally distributed variables μ_{ij} represent the mean of the Y_{ij}, ${\sigma_j}^2$ is the variance of the Y_{ij} variable, β_{0j}, β_{lj} parameters of the regression of μ_{ij} on latent variables Z_{li} , $D_l(\alpha_l, G_{lo})$is the Dirichlet

process mixture over latent variable *l* with precision parameter α_l and centring distribution G_{lo}, where G_{lo} is normally distributed with mean θ_{li} and variance V_{li}.

7.5 Priors, result determination and convergence

Priors were kept the same as in the previous chapter on correlated variables for all latent variables and parameters, see chapter 6. Results for the number of mixtures, multimodality statistics and the percentage of correct membership were determined in the same way as in the previous chapter on correlated outcomes. Convergence was established as a percentage and reported as such in the same way as the correlated outcome simulations.

7.6 Estimation of the number of factors

To determine the correct number of factors in a Bayesian factor analysis is difficult as the amount of variance accounted for each factor cannot be computed in the same way as a classic factor analysis. Usual model criteria, such as DIC do not work as these tend to decrease with the number of factors added resulting in a large number of factors and an over-fitted model. Recent papers have used reversible jump MCMC (Viroli,C, 2009) to jump between models with different numbers of factors but these add an extra layer of complicity to an already long-running and complicated model.

For these reasons the factors were determined a prior by using a normally distributed factor analysis model and applying the Kaiser criteria to determine the number of factors to use. The Kaiser criterion determines the number of factors by only keeping the ones that have an eigenvalue greater than 1. Once these are chosen the same number of factors are applied in the DPNMLM with the variable with the highest factor loading acting as an anchor for that factor.

7.7 Scenarios

7.7.1 Scenario 1 two correlated factors

The two factors used to create the underlying distribution are both correlated with covariance matrix = $\begin{matrix} 1 & 0.2 \\ 0.2 & 1 \end{matrix}$. The first factor is composed of two mixtures in a 1:4 ratio with the first bigger mixture having mean -5 and the second smaller mixture having mean 5 both with variance equal to 1, see graph 7.1.1.1, while the second factor is composed of a single normal distribution with mean 20 and variance 1, see graph 7.1.1.4. The parameters used to simulate the variables are chosen in table 7.7.1.1. The variables were first used in a classic factor analysis to determine the number of factors with eigenvalue greater than 1, see table 7.1.1.2 and to determine factor anchors highlighted in yellow, see table 7.1.1.3. Graph 7.1.1.2 and 7.1.1.5 are the posterior distribution of factors 1 and 2 respectively and graph 7.1.13 and 7.1.1.6 are the dendrograms associated with factors 1 and 2 respectively. The dip statistics and percentage membership can be seen in table 7.1.1.4.

Table 7.7.1.1: Parameters chosen for simulations in latent variable model

Variable Number Y_j	β_0	β_1	β_2	σ
Y_1	5	3	0	1
Y_2	2	0	2	3
Y_3	2	0	1	1
Y_4	25	2	0	2
Y_5	4	0	3	2
Y_6	100	1	0	2
Y_7	75	3	0	2
Y_8	2	1	0	3
Y_9	10	0	1	1

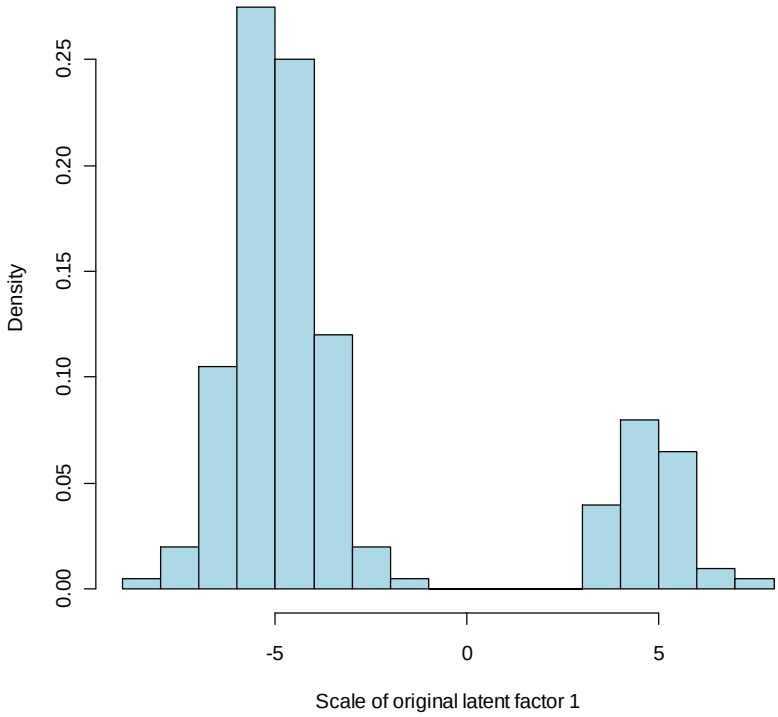
Table 7.7.1.2 normally distributed classic factor analysis eigenvalues

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.823	42.482	42.482
2	2.767	30.741	73.223
3	.930	10.334	83.557
4	.687	7.634	91.191
5	.449	4.987	96.177
6	.154	1.716	97.894
7	.094	1.047	98.941
8	.058	.646	99.587
9	.037	.413	100.000

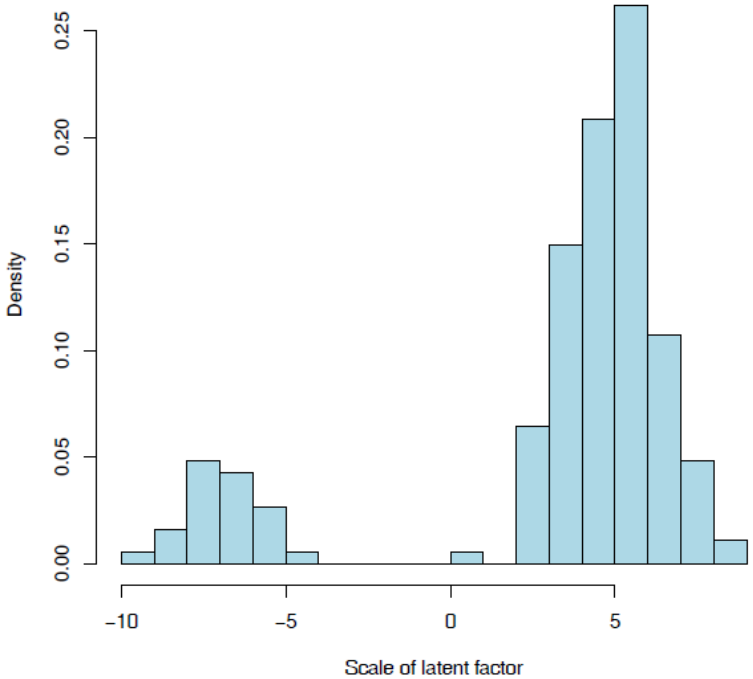
Table 7.7.1.3 Factor loading with variable factor anchors (highlighted) to be used in the Dirichlet Process Normal Mixture latent variable model

	Factors	
	1	2
y1	.069	.915
y4	.071	.760
y3	.976	-.068
y2	.961	-.037
y5	.983	-.055
y6	.059	.673
y7	.034	.887
y8	.149	.314
y9	.972	-.080

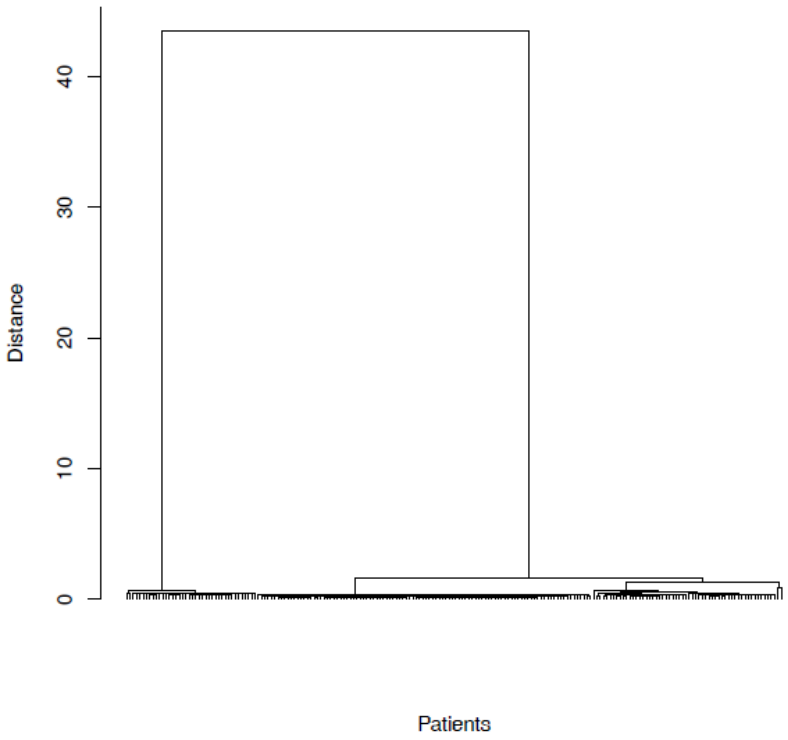
Graph 7.7.1.1 Original distribution of factor 1



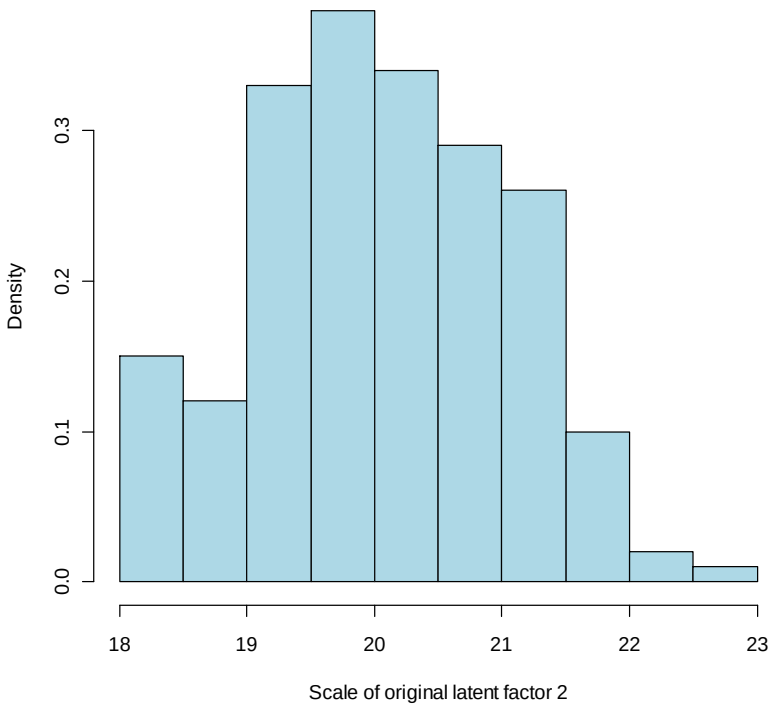
Graph 7.7.1.2 Posterior to show distribution of factor 1



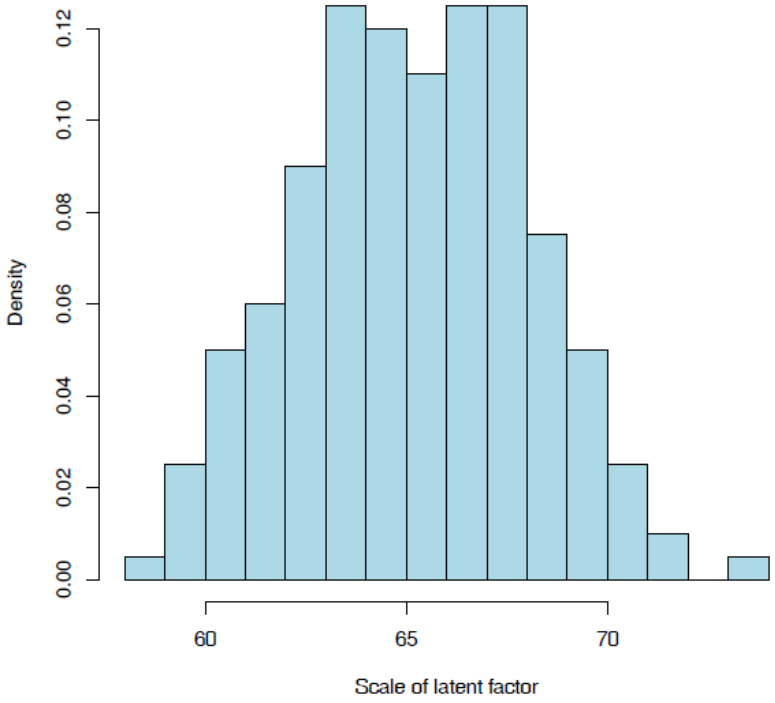
Graph 7.7.1.3 Dendrogram of factor 1



Graph 7.7.1.4 Original distribution of factor 2



Graph 7.7.1.5 Posterior to show distribution of factor 2



Graph 7.7.1.6 Dendrogram of factor 2

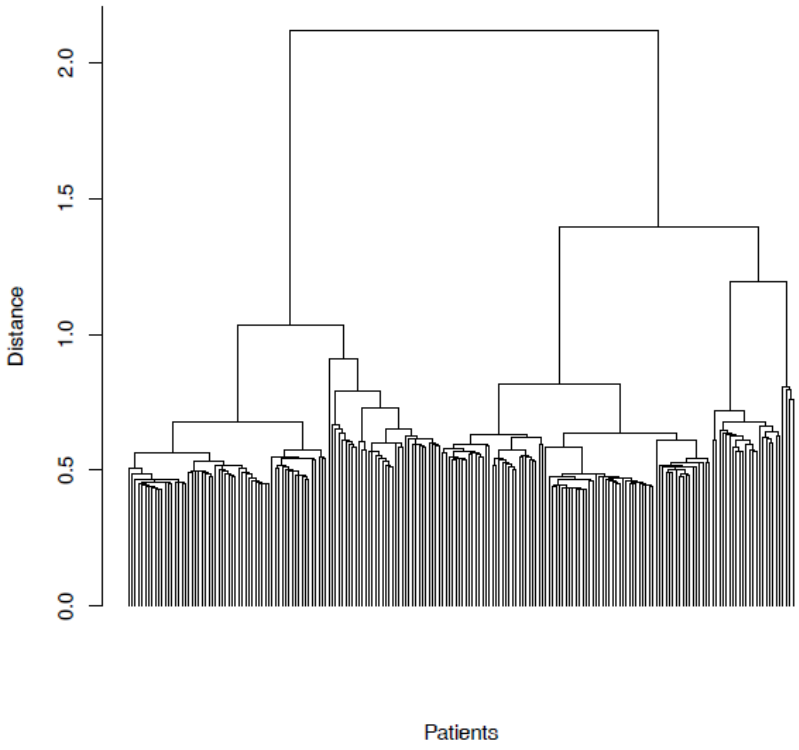


Table 7.7.1.4 Statistics taken from factor 1 and 2

Factor Number	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	2	100	100.00%	0.03957
2	2	NA	66.94%	0.02063

7.7.2 Scenario 2, three correlated factors

The three factors used to create the underlying distribution are all correlated with

covariance matrix = $\begin{matrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{matrix}$. The first factor is composed of a single normal distribution with mean 10 and variance 1, see graph 7.7.2.1 while the second factor is comprised of two mixtures in a 1:4 ratio with the first bigger mixture having mean -5 and the second smaller mixture having mean 5 both with variance equal to 1, see graph 7.7.2.4. The third factor is also comprised of two mixtures in a 1:4 ratio with the smaller mixture having a mean of -125 and the larger mixture having a mean of -20 both with variance equal to 1, see graph 7.7.2.7. The parameters used to simulate the variables are chosen in table 5. The variables were first used in a classic factor analysis to determine the number of factors with eigenvalue greater than 1, see table 7.7.2.2 and to determine factor anchors highlighted in yellow, see table 7.7.2.3. Graph 7.7.2.2

and 7.7.2.5 are the posterior distribution of factors 1 and 2 respectively and graph 7.7.2.3 and 7.7.2.6 are the dendrograms associated with factors 1 and 2 respectively. The dip statistics and percentage membership can be seen in table 7.7.2.4.

Table 7.7.2.1 Parameters chosen for simulations in latent variable model

Variable Number Y_j	β_0	β_1	β_2	β_3	σ
Y_1	5	0	0	1	1
Y_2	2	0	2	0	3
Y_3	2	0	0	2	1
Y_4	25	2	0	0	2
Y_5	4	0	3	0	2
Y_6	100	1	0	0	2
Y_7	75	3	0	0	2
Y_8	2	0	0	3	3
Y_9	10	0	1	0	1

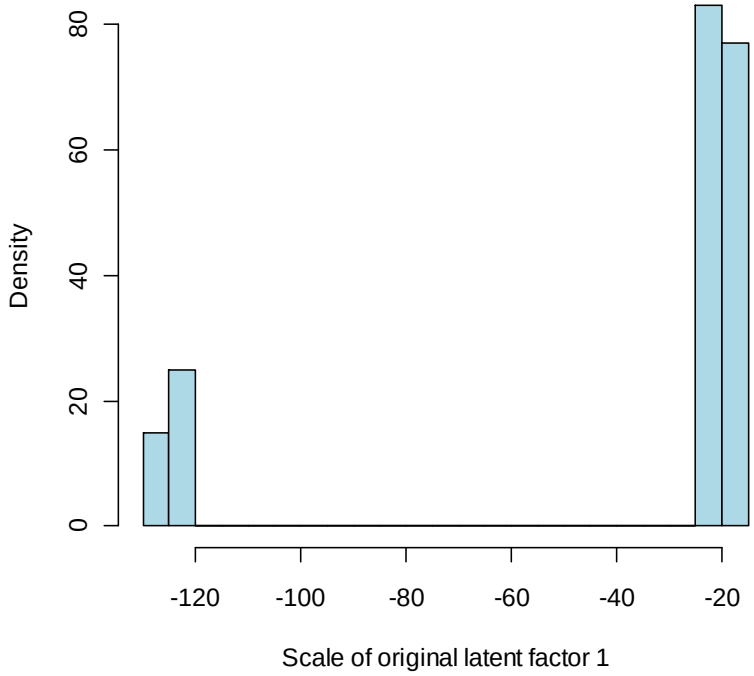
Table 7.7.2.2 normally distributed classic factor analysis eigenvalues

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.759	63.984	63.984
2	1.844	20.494	84.479
3	.747	8.303	92.782
4	.402	4.470	97.252
5	.138	1.534	98.786
6	.072	.800	99.585
7	.036	.406	99.991
8	.001	.006	99.997
9	.000	.003	100.000

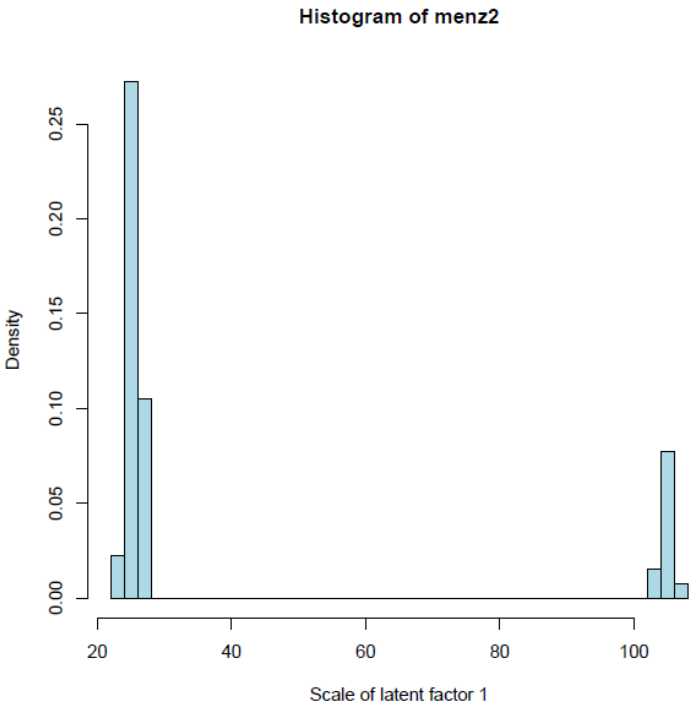
Table 7.7.2.3 Factor loading with variable factor anchors (highlighted) to be used in Dirichlet Normal mixture process model

	Factors	
	1	2
y1	.989	.010
y4	-.947	.047
y3	.990	.013
y2	.023	.826
y5	-.979	.057
y6	.141	.639
y7	.034	.862
y8	.989	.009
y9	-.972	.072

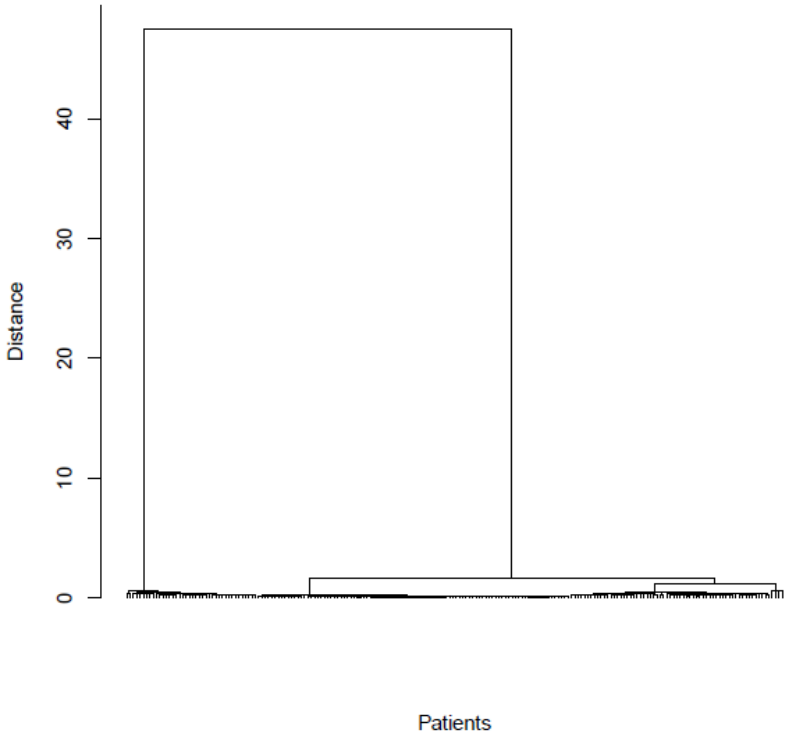
Graph 7.7.2.1 Original distribution of factor 1



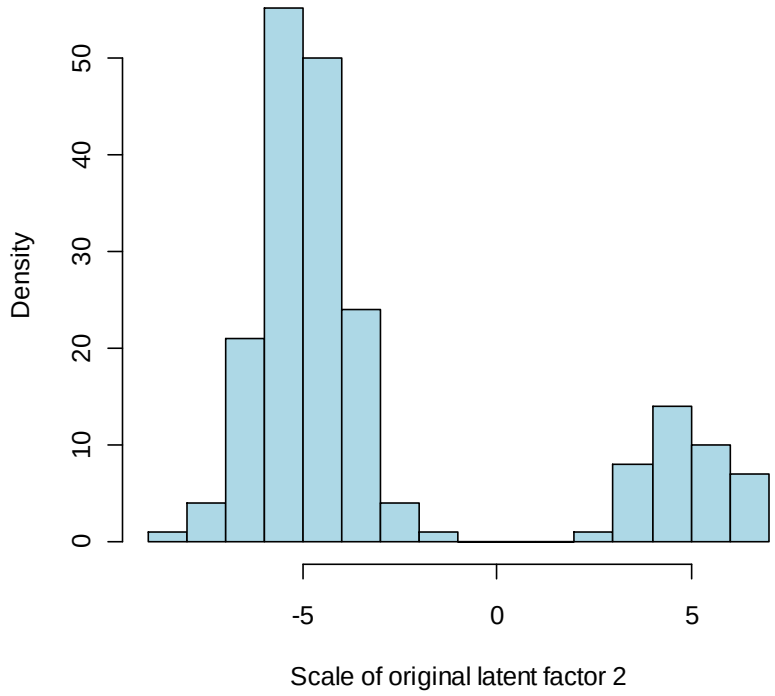
Graph 7.7.2.2 Posterior to show distribution of factor 1



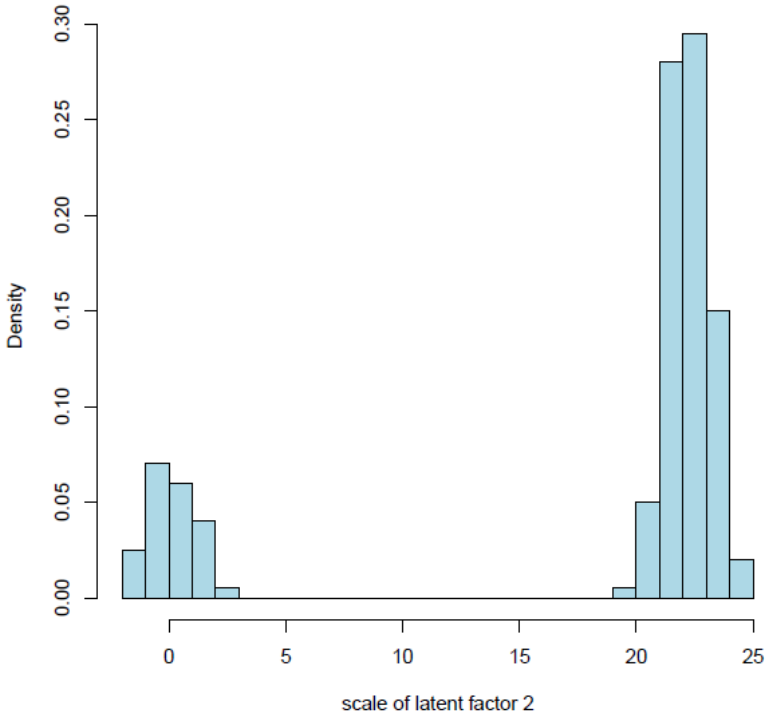
Graph 7.7.2.3 Dendrogram of factor 1



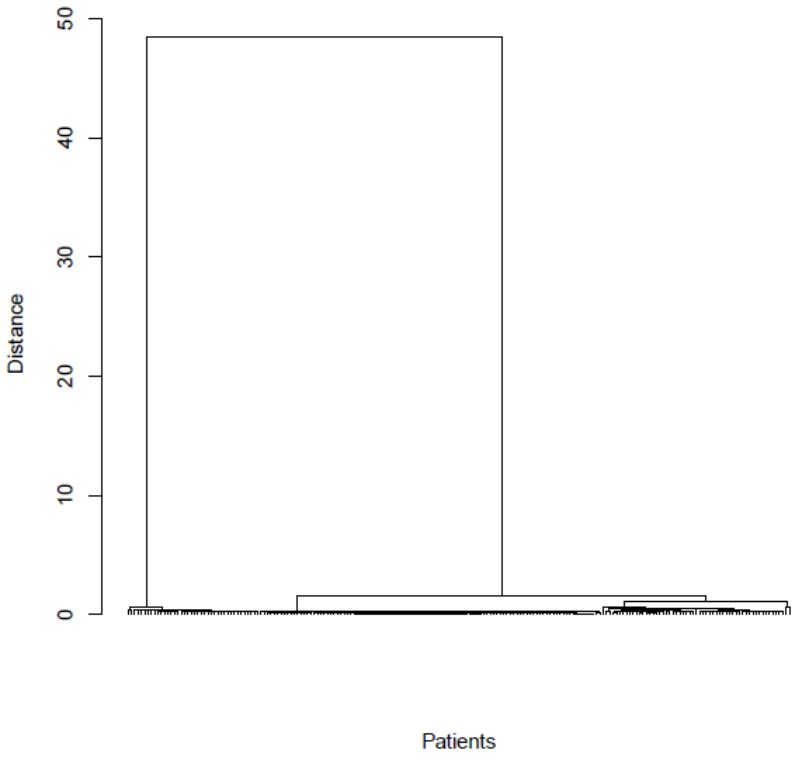
Graph 7.7.2.4 Original distribution of factor 2



Graph 7.7.2.5 Posterior to show distribution of factor 2



Graph 7.7.2.6 Dendrogram of factor 2



Graph 7.7.2.7 Original distribution of factor 3

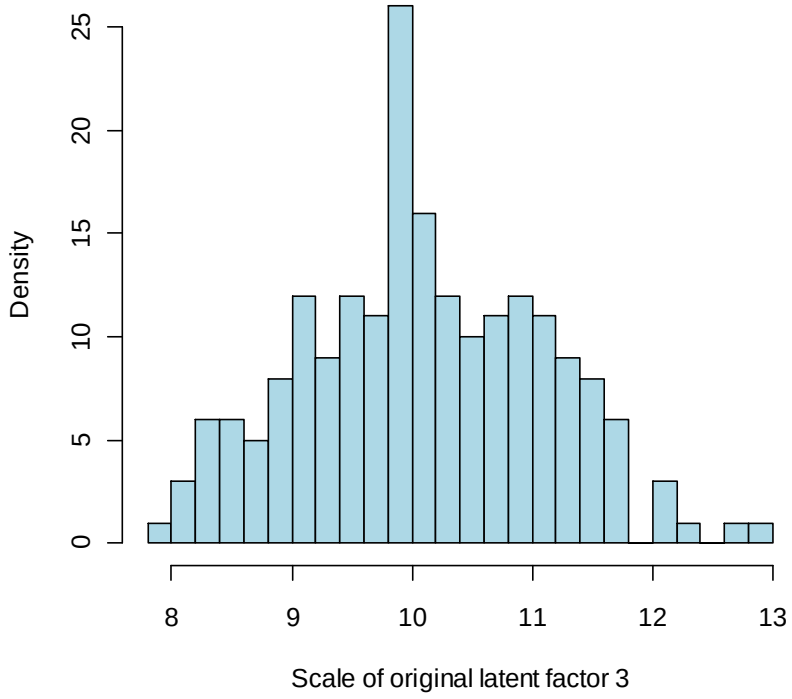


Table 7.7.2.4 Statistics taken from factor 1 and 2

Factor Number	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1		NA	100%	0.0690
2	2	NA	100%	0.0805

7.7.3 Scenario 3 two uncorrelated factor

The two factors used to create the underlying distribution were both uncorrelated. The first factor is composed of a mixture of two normal distributions in a ratio of 1:2 with mean 5 and variance 1 for the smaller mixture and mean=0 and variance =1 for the larger mixture, see graph 7.7.3.1. The second factor is composed of two mixtures in a 1:9 ratio with the first bigger mixture having mean 0 and the second smaller mixture have mean 10 both with variance equal to 1, see graph 7.7.3.4. The parameters used

to simulate the variables are as chosen in table 7.7.3.1. The variables were first used in a classic factor analysis to determine the number of factors with eigenvalue greater than 1, see table 7.7.3.2 and to determine factor anchors highlighted in yellow, see table 7.7.3.3. Graph 7.7.3.2 and 7.7.3.5 are the posterior distribution of factors 1 and 2 respectively and graphs 7.7.3.3 and 7.7.3.4 are the dendrograms associated with factors 1 and 2 respectively. The dip statistics and percentage cluster membership can be seen in table 7.7.3.4.

Table 7.7.3.1 Parameters chosen for simulations in latent variable model

Variable Number Y_j	β_0	β_1	β_2	σ
Y_1	5	3	0	1
Y_2	2	0	2	3
Y_3	2	0	1	1
Y_4	25	2	0	2
Y_5	4	0	3	2
Y_6	100	1	0	2
Y_7	75	3	0	2
Y_8	2	1	0	3
Y_9	10	0	1	1

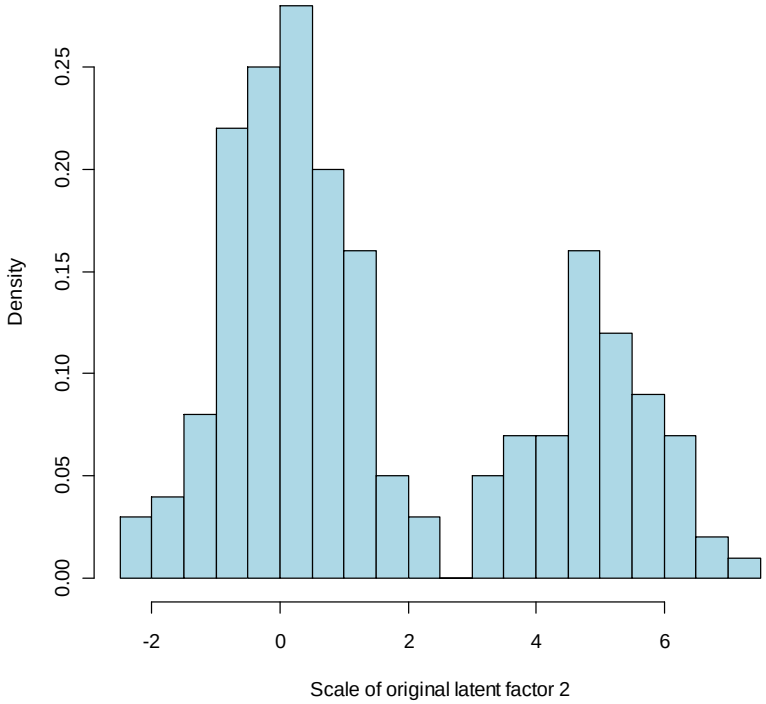
7.7.3.2 Normally distributed classic factor analysis eigenvalues

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.427	60.297	60.297
2	2.268	25.205	85.502
3	.495	5.502	91.005
4	.363	4.035	95.040
5	.184	2.049	97.088
6	.089	.984	98.073
7	.087	.963	99.036
8	.051	.567	99.603
9	.036	.397	100.000

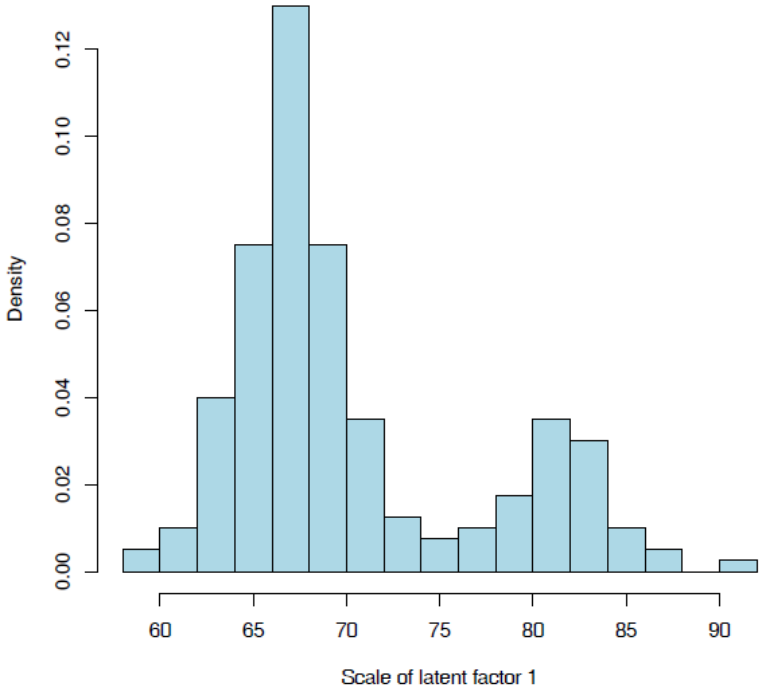
Table 7.7.3.3 Factor loading with variable factor anchors (highlighted) to be used in Dirichlet Normal mixture process model

	Factors	
	1	2
y1	.845	.475
y4	.757	-.536
y3	.765	-.578
y2	.820	.466
y5	.806	-.555
y6	.696	.448
y7	.849	.448
y8	.648	.410
y9	.779	-.573

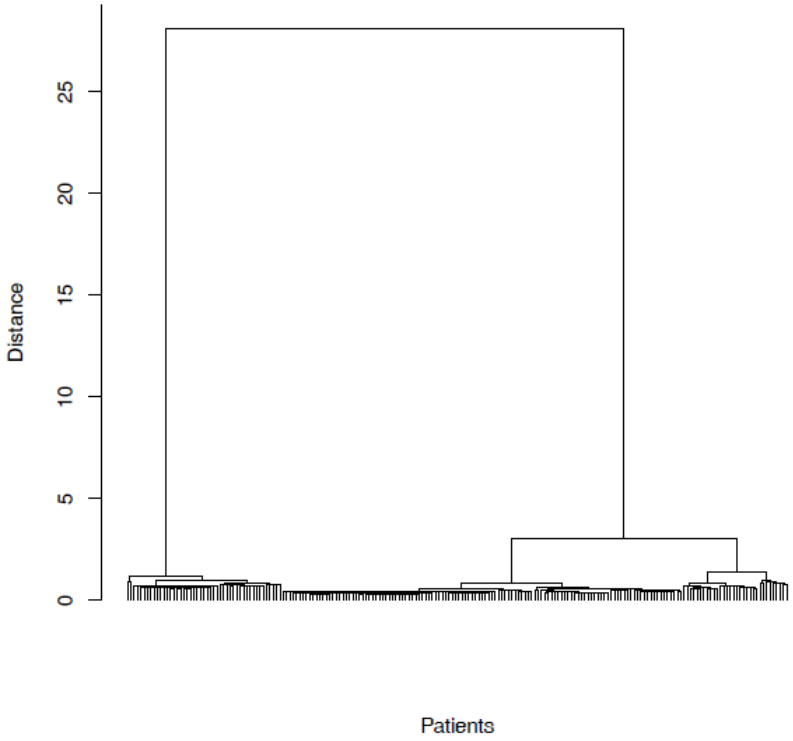
Graph 7.7.3.1 Original distribution of factor 1



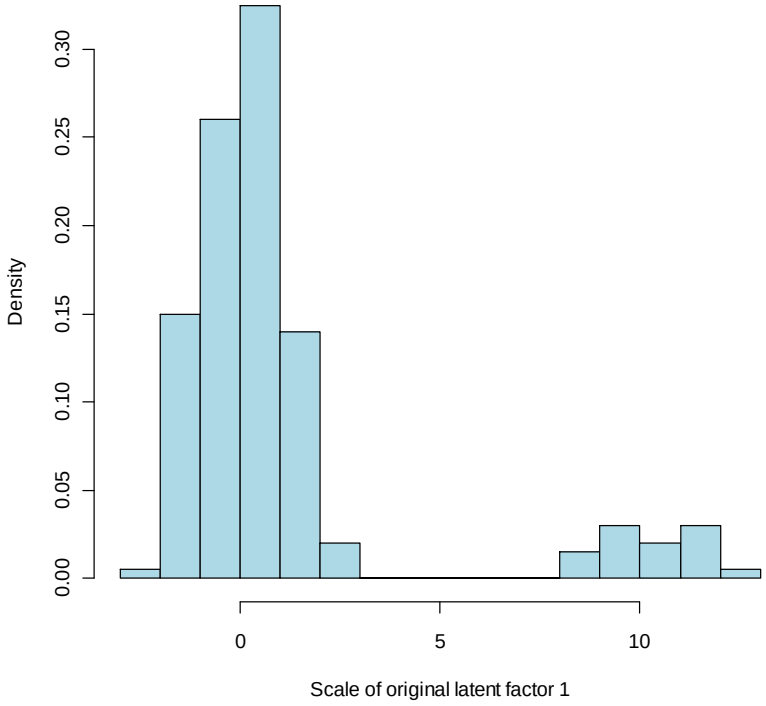
Graph 7.7.3.2 Posterior to show distribution of factor 1



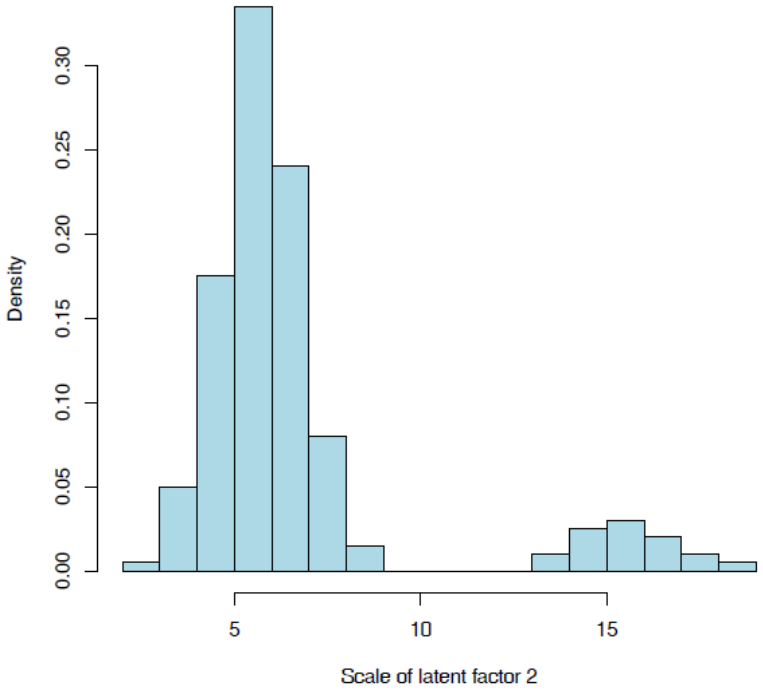
Graph 7.7.3.3 Dendrogram of factor 1



Graph 7.7.3.4 Original distribution of factor 2



Graph 7.7.3.5 Posterior to show distribution of factor 2



Graph 7.7.3.6 Dendrogram of factor 2

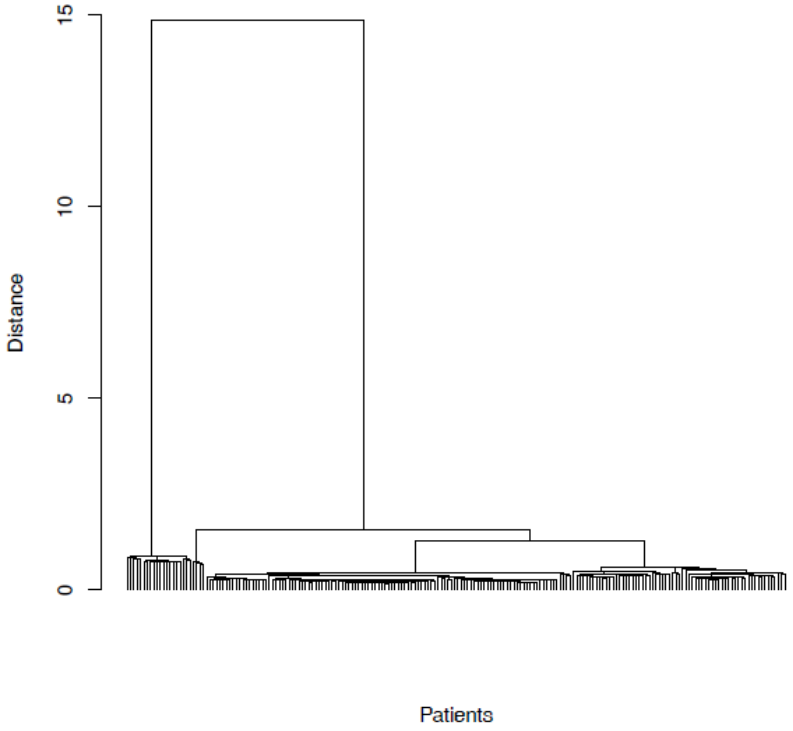


Table 7.7.3.4 The mixture and density distribution statistics from each factor

Factor Number	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	2	90%	100%	0.03084
2	2	100%	100%	0.02536

7.7.4 Scenario 4 Three uncorrelated factors

The three factors used to create the underlying distribution are uncorrelated, the first factor is normally distribution with mean 10 and variance = 2, see graph 7.7.4.1. The second factor is composed of three mixtures in a 1:1:1 ratio with the mixtures having means of 5, 0 and -5 each with unit variances, see graph 7.7.4.4, the third factor had a distribution consisting of two mixtures with a 1:9 split with the smaller mixture having mean 10 and variance 1 and the larger mixture having mean 0 and standard deviation 2, see graph 7.7.4.7. The parameters used to simulate the variables are chosen in table 7.7.4.1. The variables were first used in a classic factor analysis to determine the number of factors with eigenvalue greater than 1, see table 7.7.4.2 and to determine factor anchors highlighted in yellow, see table 7.7.4.3. Graphs 7.7.4.2, 7.7.4.5 and 7.7.4.8 are the posterior distribution for factors 1, 2 and 3 respectively and graphs 7.7.4.3, 7.7.4.6 and 7.7.4.9 are the dendrograms associated with factors 1, 2 and 3 respectively. The dip statistics and percentage membership can be seen in table 7.7.4.5.

Table 7.7.4.1 Parameters chosen for simulations in latent variable model

Variable Number Y_j	β_0	β_1	β_2	B_3	σ
Y_1	5	0	0	1	1
Y_2	2	0	2	0	3
Y_3	2	0	0	2	1
Y_4	25	2	0	0	2
Y_5	4	0	3	0	2
Y_6	100	1	0	0	2
Y_7	75	3	0	0	2
Y_8	2	0	0	3	3
Y_9	10	0	1	0	1

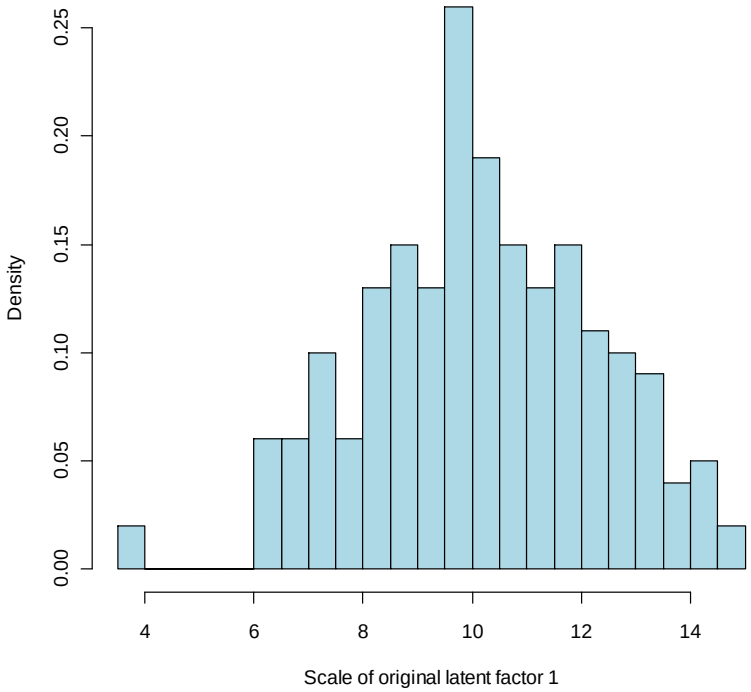
Table 7.7.4.2 Normally distributed classic factor analysis eigenvalues

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.053	45.030	45.030
2	2.470	27.445	72.475
3	1.809	20.099	92.574
4	.185	2.051	94.625
5	.156	1.737	96.362
6	.126	1.396	97.758
7	.108	1.201	98.959
8	.059	.650	99.609
9	.035	.391	100.000

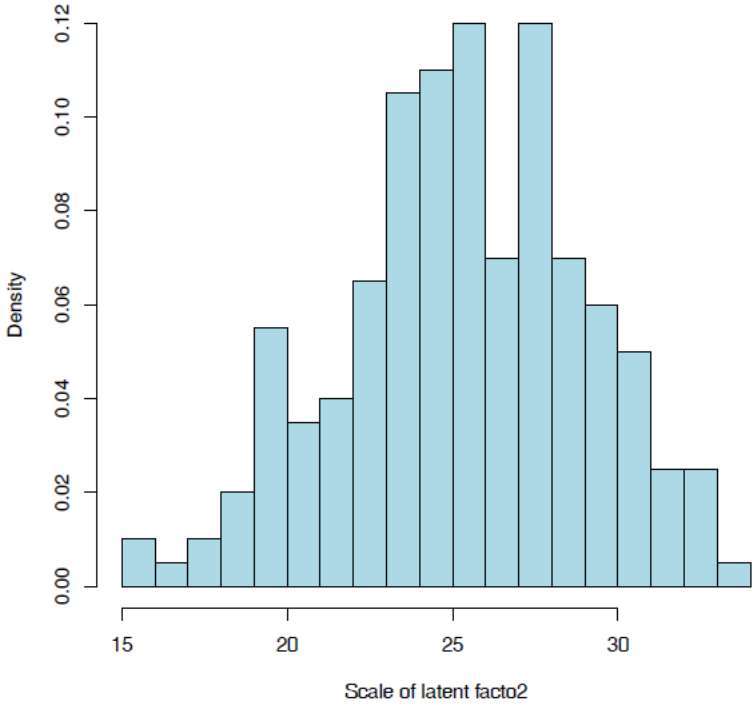
Table 7.7.4.3 Factor loading with variable factor anchors (highlighted) to be used in Dirichlet Normal mixture process model

	Factors		
	1	2	3
y1	.449	.830	.048
y4	.459	.845	.031
y3	.780	-.221	-.547
y2	.722	-.198	-.583
y5	.783	-.216	-.544
y6	.760	-.297	.538
y7	.737	-.289	.537
y8	.764	-.268	.539
y9	.441	.830	.048

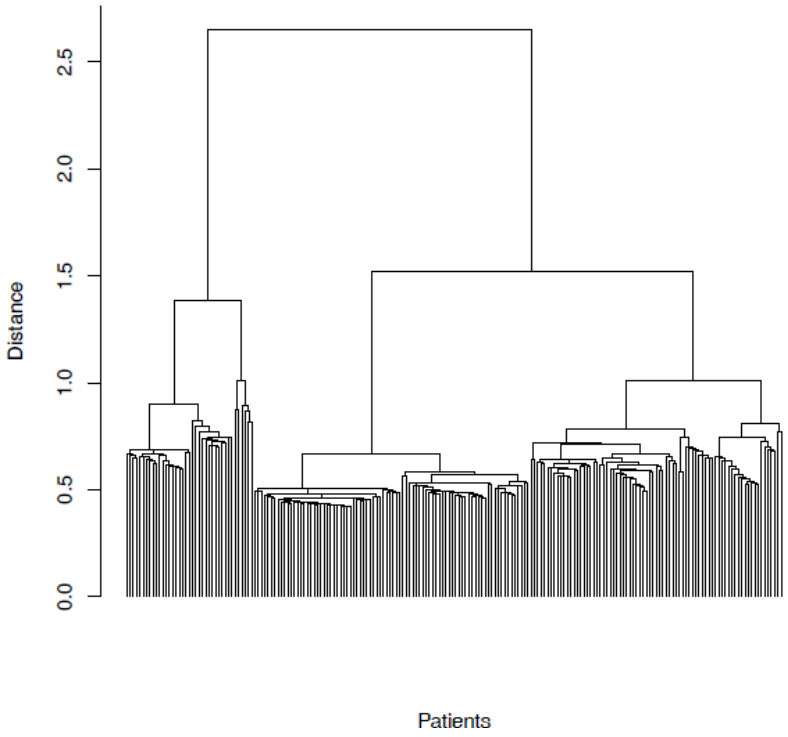
Graph 7.7.4.1 Original distribution of factor 1



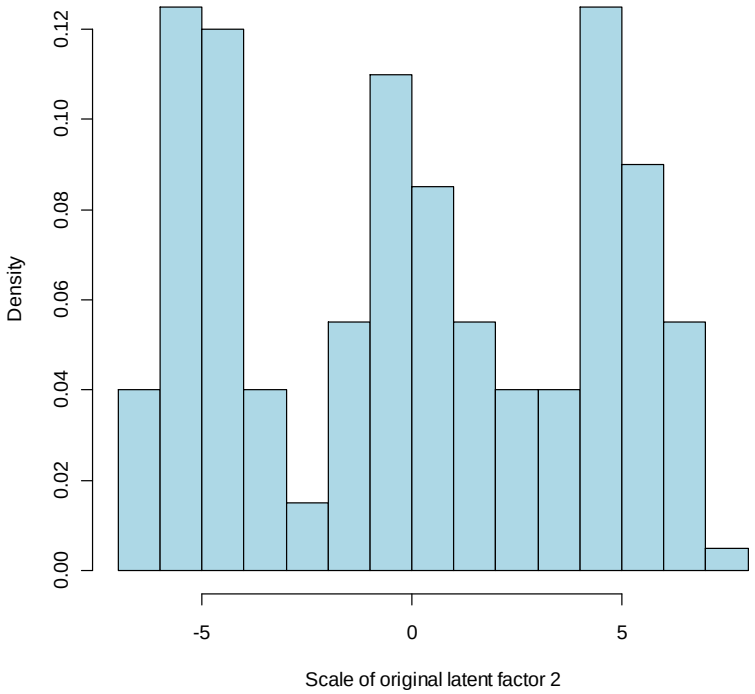
Graph 7.7.4.2 Posterior to show distribution of factor 1



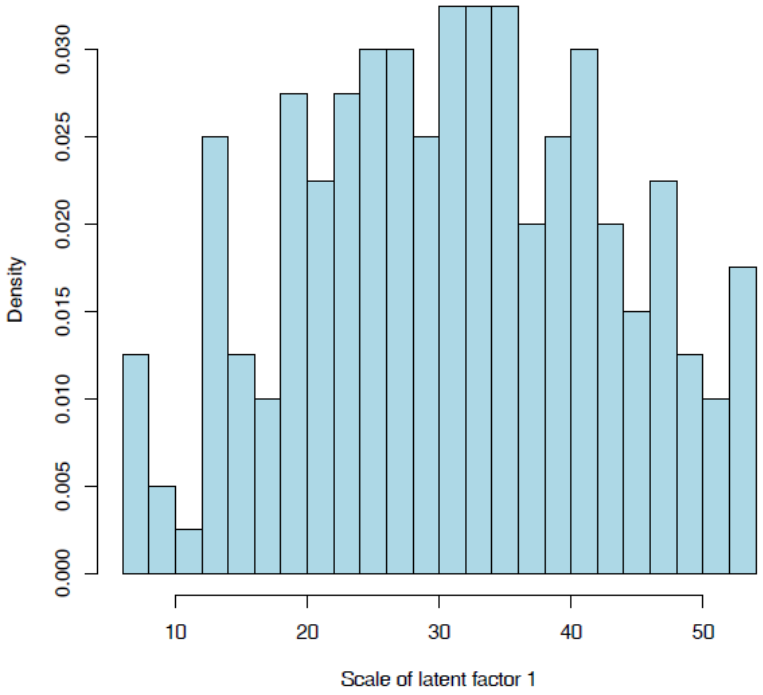
Graph 7.7.4.3 Dendrogram of factor 1



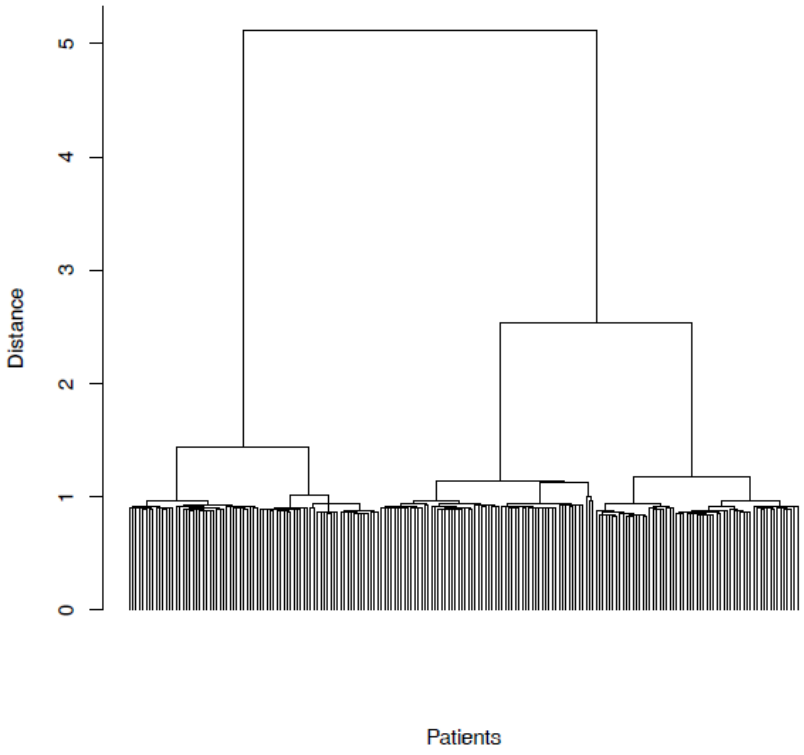
Graph 7.7.4.4 Original distribution of factor 2



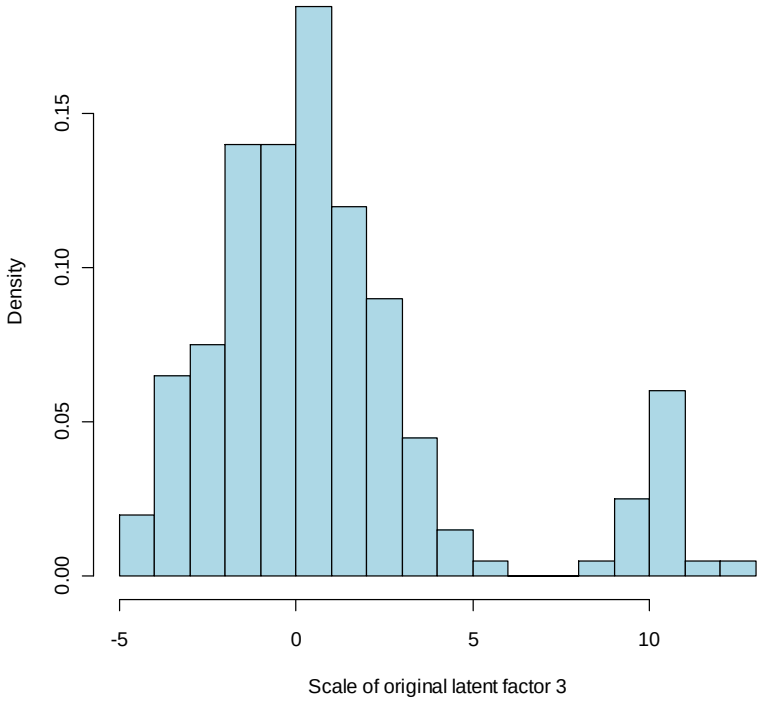
Graph 7.7.4.5 Posterior to show distribution of factor 2



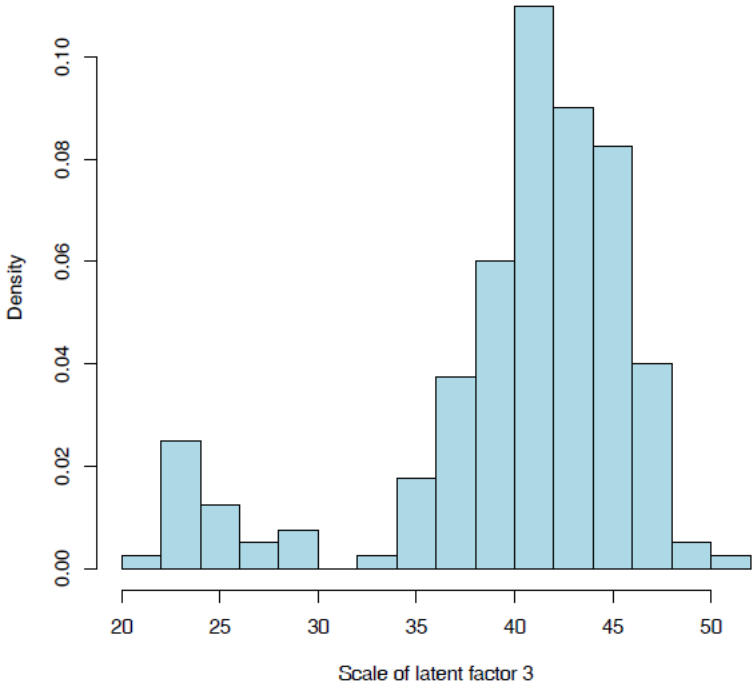
Graph 7.7.4.6 Dendrogram of factor 2



Graph 7.7.4.7 Original distribution of factor 3



Graph 7.7.4.8 Posterior to show distribution of factor 3



Graph 7.7.4.9 Dendrogram of factor 3

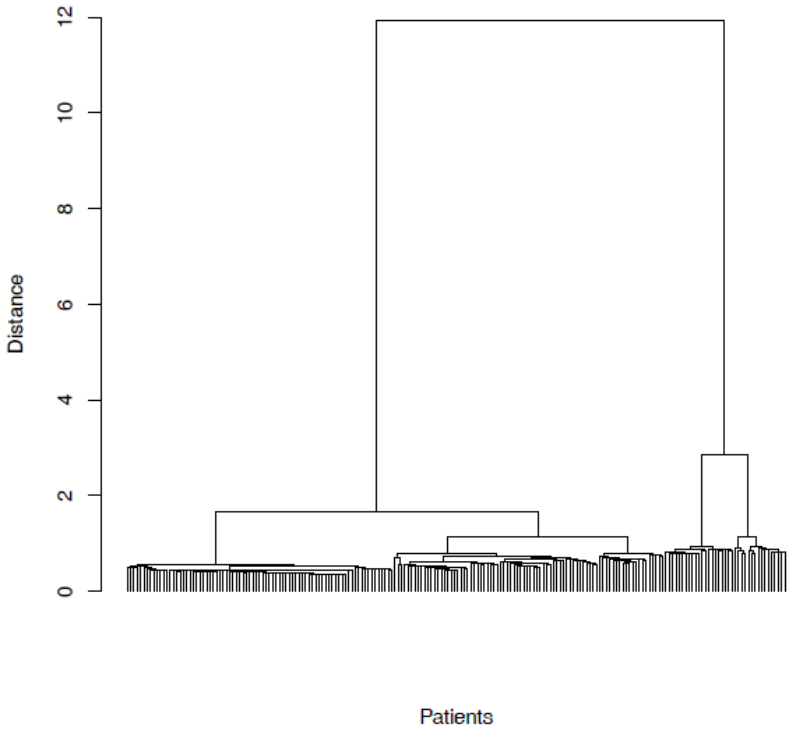


Table 7.7.4.4 The mixture and density distribution statistics from each factor

Factor Number	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	2	NA	52.11%	0.01922
2	2	NA	51.07%	0.01892
3	3	100%	90.26%	0.02202

7.7.5 Scenario 5: One factor only

The factor used to create the underlying distribution is composed of two mixtures in a 1:1 ratio with the first mixture having mean 0 and the second mixture have mean 10 both with variance equal to 1, see graph 7.7.5.1. The parameters used to simulate the variables are chosen in table 7.7.5.1. The variables were first used in a classic factor analysis to determine the number of factors with eigenvalue greater then 1, see table 7.7.5.2, and to determine factor anchors highlighted in yellow, see table 7.7.5.3. Graph 7.7.5.2 is the posterior distribution for the factor graph 7.7.5.3 is the dendrogram associated with factor 1. The dip statistics and percentage membership can be seen in table 7.7.5.4.

Table 7.7.5.1 Parameters chosen for simulations in latent variable model

Variable Number Y_j	β_0	σ
Y_1	5	1
Y_2	2	3
Y_3	2	1
Y_4	25	2
Y_5	4	2
Y_6	100	2
Y_7	75	2
Y_8	2	3
Y_9	10	1

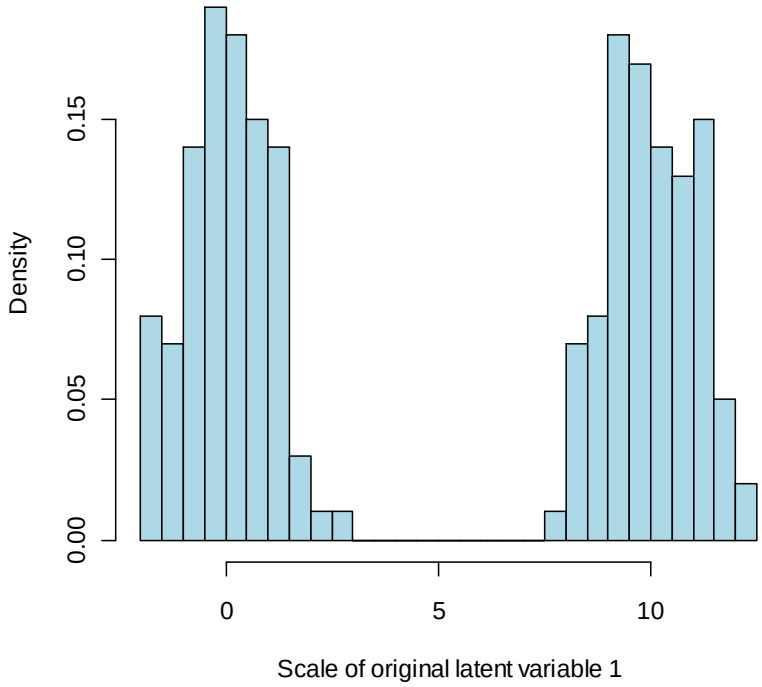
Table 7.7.5.2 Normally distributed classic factor analysis eigenvalues

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	8.440	93.774	93.774
2	.231	2.567	96.341
3	.135	1.497	97.838
4	.097	1.073	98.911
5	.036	.401	99.312
6	.030	.330	99.642
7	.017	.184	99.827
8	.010	.106	99.933
9	.006	.067	100.000

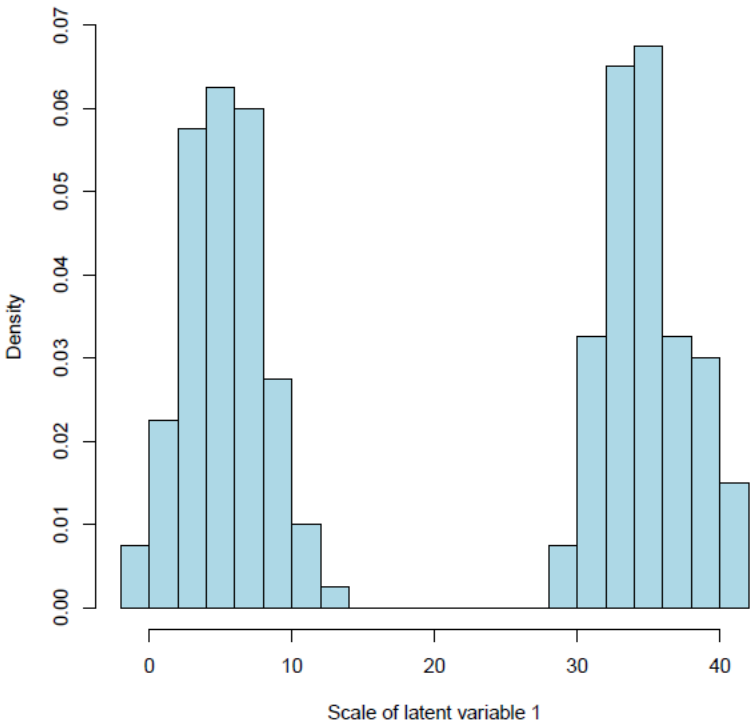
Table 7.7.5.3 Factor loading with variable factor anchors (highlighted) to be used in Dirichlet Normal mixture process model

	Factors
	1
y1	.995
y4	.890
y3	.993
y2	.982
y5	.990
y6	.950
y7	.981
y8	.989
y9	.940

Graph 7.7.5.1 Original distribution of factor 1



Graph 7.7.5.2 Posterior to show distribution of factor 1



Graph 7.7.5.3 Dendrogram of factor 1

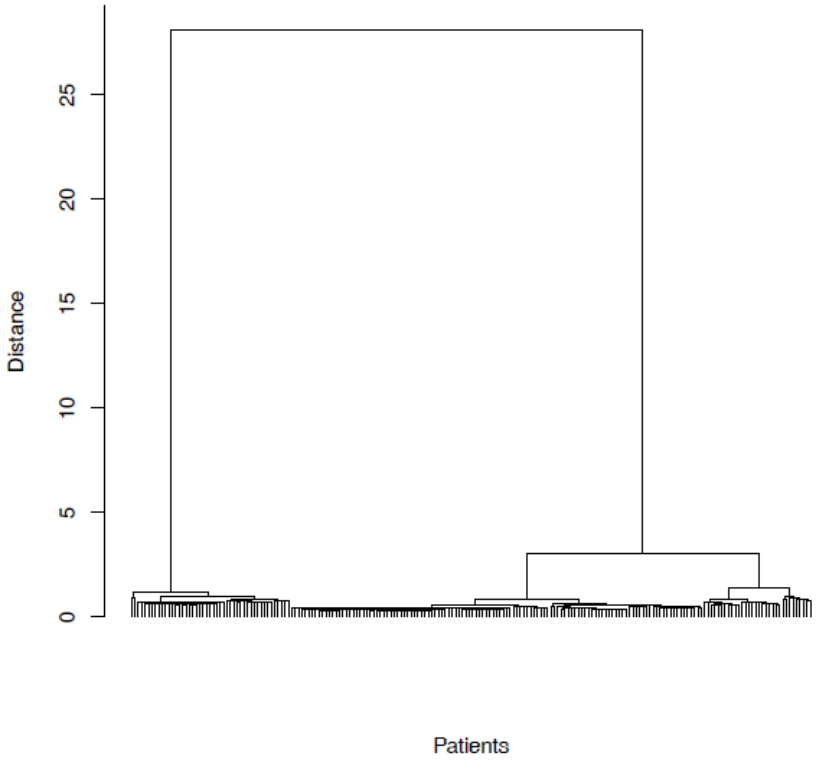


Table 7.7.5.4 The mixture and density distribution statistics from the factor

Factor Number	Number of mixtures found for maximum distance in dendrogram	Percentage correct cluster membership	Percentage of iterations accepting null Hypothesis of Diptest (Multimodality) at p=0.05	Mean of dip statistic for all iterations
1	2	100%	100%	0.1495

7.8 Conclusion

The Dirichlet process normal mixture latent variable model works well for determining mixtures over factors, when the factor number have been pre-specified a prior. The correct number of mixtures is returned for well defined distributions that have visible sub-groups and good percentage membership is also returned. For poorer defined sub-groups as is the case in scenario 4 factor 2 the mixtures returned reflects the fact that there is possibly many ways of sampling a distribution and the model returns the simplest explanation that being 2 groups instead of three. Even when these two groups were returned the cluster membership resembled that of the three mixtures with the first posterior subgroup being comprised solely of group 1 from the original distribution and posterior group 2 containing a mixture of the original group 2 and 3. The mean dip statistic gave a good indication if sub-groups were present with the dip statistics being greater than p=0.05 cut off of 0.01897, although it failed to detect uni-modal distributions as these were just over the p=0.05 cut off.

Deciding the number of factors a prior using a classic factor analysis is a good way of determining the correct number of factors to use in a Dirichlet process normal mixture latent variable model. Only one of the scenarios returned a different number of factors then the ones simulated this was the case in scenario three. In scenario three, 3 correlated factors were simulated but the classical analysis only detected two this is due to the factors being slightly correlated in the simulation, but if independent factors are needed then we can obtain these and it is an acceptable assumption to only model independent factors for these complex models (Viroli, C, 2009).

7.9 Chapter closing statement

The methodology tested here for the Dirichlet Process Normal Mixture Latent variable Models (DPNMLVM) for multiple factors and correlated/uncorrelated outcomes can now be used to determine mixtures over factors for real data and return correct results for independent factors over a varying amount of mixture distributions. This methodology will now be applied to two severe asthma datasets in the next two chapters, chapter 8, analysis of the Haldar dataset and chapter 9, analysis of the Brompton blood dataset.

Chapter 8. Analysis on Brompton Blood Dataset

8.1 Chapter outline

The Brompton blood dataset is now analysed in this chapter using the Dirichlet process normal mixture latent variable model, DPNMLVM 4 factors are found with clusters over those factors but the clusters were not significantly multimodal suggesting that the distributions are not genuine clusters but make up a larger non normal distribution. The specificity analysis is carried out and this indicated that the prior on alpha especially when $\alpha=1$ did make a difference to the clustering possible due to the small amount of data, leading to the conclusion that the dataset is underpowered to detect significant clusters.

8.2 Introduction to the dataset

The Brompton Blood dataset was originally created for genotyping a small number of severe asthma patients and contains several phenotypic outcomes that were taken from patients attending the severe asthma clinic at the Glenfield Hospital, Leicester. The dataset contains demographics and asthma related biomarkers for 157 patients, as are summarised in chapter 3. The final dataset used was a subset containing 120 patients with no missing data.

8.3 Variables

The outcomes used in the Dirichlet Process Normal Mixture Latent Variable Model (DPNMLVM) are Total blood IgE which describes a blood count of IgE antibodies a high amount of these are related to an atopic response, Body Mass Index (BMI), JACS score an asthma symptom questionnaire, sputum inflammation cell counts logs for

neutrophils and eosinophils. Also there were four variables relating to spirometry these were Pre and post bronchodilator Forced Expired Volume in one second (FEV1)/ predicted FEV1 which measures air flow obstruction allowing for gender, age and height differences and Forced Volume Capacity, (FVC) a measure of the volume of the lungs. These variables were chosen to be similar to the previous analysis to see if similar clusters could be obtained, in addition the spirometry measurements were added as these are usually used in nearly all analyses of asthma and respiratory disease as they are comparable with breathlessness, the main symptom of asthma.

8.4 Classic factor analysis

Factor analysis was carried out on the variables under the usual assumption of normally distributed latent/factor variables to determine the number of factors and to determine which variables would be factor anchors in the DPNMLVM as in the Haldar dataset. Four factors were found to be sufficient using the Kaiser criteria to determine the number of factors. The four factors explained around 77% of the variance seen in the dataset. See Table 8.3.1. The four factor anchors to be used in the DPNMLVM were chosen as in the previous Haldar factor analysis by selecting the variables that had the highest factor loading on a factor, See table 8.3.2.

Table 8.4.1: The results of the standard factor analysis of the 9 variables from the Brompton severe asthma dataset. A four factor solution satisfies the Kaiser criteria and allows for 77% of the variance.

No of Factors	Eigenvalues	Cumulative % of Variance
1	3.043	33.813
2	1.493	50.407
3	1.281	64.645
4	1.116	77.049
5	0.870	86.717
6	0.602	93.403
7	0.435	98.235
8	0.131	99.689
9	0.028	100.000

Table 8.4.2: Shows the factor loading for the four normal factor model, the highest factor loading for each variable has been highlighted and used for annotation of the factors. The highest loading variable for each factor was used as a factor anchor for that factor in the Dirichlet Process Normal Mixture Latent variable Model (DPNMLVM).

Variables	1	2	3	4
Total IgE, kU/l	.031	-.029	.631	-.082
BMI	-.164	.087	-.320	.811
JACS symptom score Mean score	-.556	.212	.253	.472
Pre FVC, L	.840	-.016	.420	.214
Pre FEV1/ predicted, %	.830	.248	-.331	-.053
Post FVC,L	.797	.008	.465	.253
Post FEV1/ Predicted,%	.730	.379	-.401	-.063
Sputum Neutrophil count, Log %	-.365	.685	.211	-.322
Log of Sputum Eosinophil count, Log %	.106	-.875	-.096	-.098

The factors can be described by the variables that they correlate to. Thus factor 1 describes a spirometry or airflow obstruction component to the dataset. Factor 2 describes an eosinophilic inflammation factor. Factor 3 describes IgE levels and thus

atopic status and factor 4 describes the role of BMI in the dataset. JACS Symptoms were negatively associated with spirometry measurements.

The anchors for each factor to be used in the Dirichlet Process Normal Mixture Latent variable Model are the highest factor loading variable for each factor these are pre bronchodilator FVC, log eosinophil cell count, Total IgE count and BMI. This is the rationale for using four factors with the four factor anchors in the DPNMLM.

8.5 Dirichlet Process Normal Mixture Latent Variable Model

For variables j for j=1,2,…9

For subjects i for i=1,2...120

For latent variable l for 1,2...4

$$Y_{ij} \sim N\left(\mu_{ij}, {\sigma_j}^2\right) \qquad \textit{Equation 206}$$

$$\mu_{ij} = \beta_{0j} + \beta_{lj}.Z_{li} \qquad \textit{Equation 207}$$

$$Z_{li} \sim D_l(\alpha_l, G_{lo}) \qquad \textit{Equation 208}$$

$$G_{lo} \sim N(\theta_{li}, V_{li}) \qquad \textit{Equation 209}$$

Where Y_{ij} represents the i individual of the j normally distributed variables μ_{ij} represent the mean of the Y_{ij}, ${\sigma_j}^2$ is the variance of the Y_{ij} variable, β_{0j}, β_{lj}

parameters of the regression of μ_{ij} on latent variables Z_{li} , $D_l(\alpha_l, G_{lo})$is the Dirichlet process mixture over latent variable l with precision parameter c_{lo} and centring distribution G_{lo}, where G_{lo} is normally distributed with mean θ_{li} and variance V_{li}. $\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}$ all equal 0 and $\beta_{11}, \beta_{22}, \beta_{33}, \beta_{44}$ all equal 1 as these are the parameters associated with the factor anchors and kept constant for identity purposes, where

Y_{i1} =Pre Bronchodilator FVC

Y_{i2} =Log eosinophil cell count

Y_{i3} =Total IgE blood count

Y_{i4} =BMI

Y_{i5} =Post Bronchodilator FVC

Y_{i6} =Pre Bronchodilator FEV1/Predicted

Y_{i7} = Post Bronchodilator FEV1/Predicted

Y_{i8} =JACS symptom score

Y_{i9} =Neutrophil cell count

8.6 Priors and Convergence

Priors were kept the same as the previous simulations, with a sensitivity analysis carried out on the alpha parameter of the model. The model parameters were tested for convergence using the Heidelberg test after 425,000 iterations for a further 75,000 after and the percentage of parameter that passed this test were 96.43%. The parameters that did not pass this test were plotted and checked for convergence.

8.7 Results

8.7.1 Variable analysis on factors

The correlation for each variable against the four factors of the DPNMLVM can be seen in table 8.6.1 this is similar to the factor loading table from table 8.4.2 for the normally distributed factor analysis, suggesting that the factors and the analysis is similar. Factor 1 corresponded to spirometry and lung volume, factor 2 to eosinophilic inflammation, factor 3 to airflow obstruction, symptoms and neutrophilic inflammation and IgE and factor 4 to BMI. With the factors established we now turn our attention to the clusters in each factor.

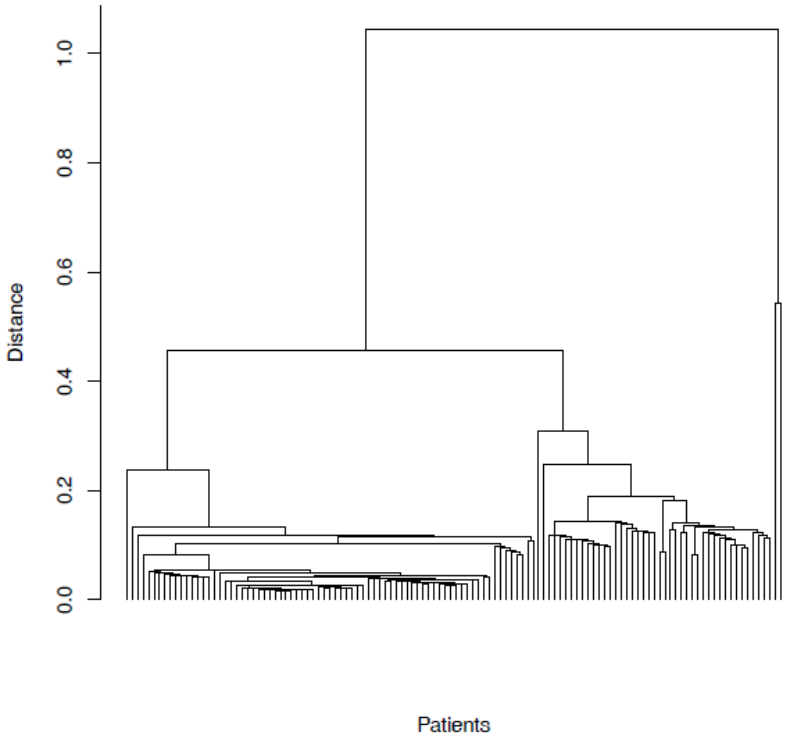
Table 8.7.1 describes the correlation of the variables with the factors in the Dirichlet Process Normal Mixture Latent variable Model

Variables	1	2	3	4
Total. IgE Count, kU/l	0.098	-0.004	-0.109	-0.075
BMI, kg/m^2	-0.169	-0.068	-0.10	0.992
JACS symptom score, mean score	-0.293	-0.199	-0.353	0.159
Pre FVC, L	0.992	-0.034	-0.091	-0.013
Pre FEV1/ Predicted, %	0.545	-0.124	0.804	0.048
Post FVC, L	0.977	-0.058	-0.161	0.005
Post FEV1/ Predicted, %	0.422	-0.206	0.761	0.059
Log Sputum Neutrophil cell count, Log %	-0.152	-0.108	-0.451	-0.218
Log Eosinophil cell count, Log %	0.004	0.995	0.027	-0.067

8.7.2 Factor 1 Lung Volume

The first factor had cluster dendrogram as in graph8.7.2.1 suggesting the factor is separated into two or possibly three clusters. Inspection of the density distribution for factor 1, See graph 8.7.2.2, suggests that the distribution is not multi-modal indicating that the 2 groups found represent a bigger non normal distribution. This is further backed up by the distribution passing the dip test of multi modality as we reject the null hypothesis of the distribution being multimodal with dip statistic being non-significant for this factor density distribution, dip statistic = 0.02370, (p=0.02). The clusters found are significant for the variables log of eosinophil count, LABA (use of long acting beta-agonist, and exacerbations, which are described as number of visits to hospital see table 8.7.2.1 The data in factor 1 is split into cluster 1 which describes a lower eosinophilic group with a higher percentage of users of LABA who experience fewer exacerbations and cluster 2 which indicates a high eosinophilic group who have more exacerbations and do not use LABA, see table 8.7.2.1, but these clusters are not multimodal so suggest that they are not true clusters.

Graph 8.7.2.1 Cluster dendrogram for factor 1 (Lung Volume) using the probability of being in a cluster with another patient to separate patients.



Graph 8.7.2.2 Histogram of the density of latent factor 1 (Lung Volume)

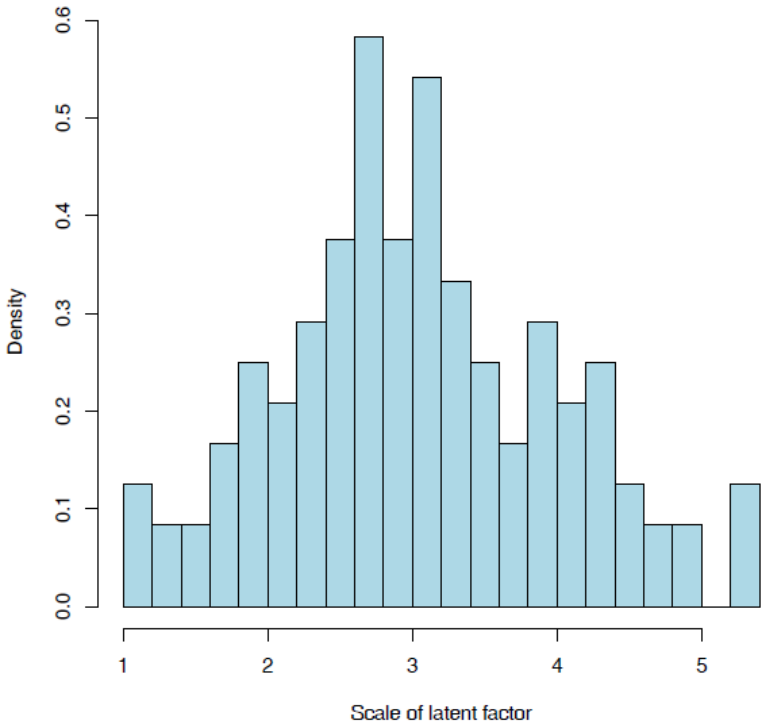


Table 8.7.2.1 The significant variables for the clusters found in factor 1 (Lung Volume)

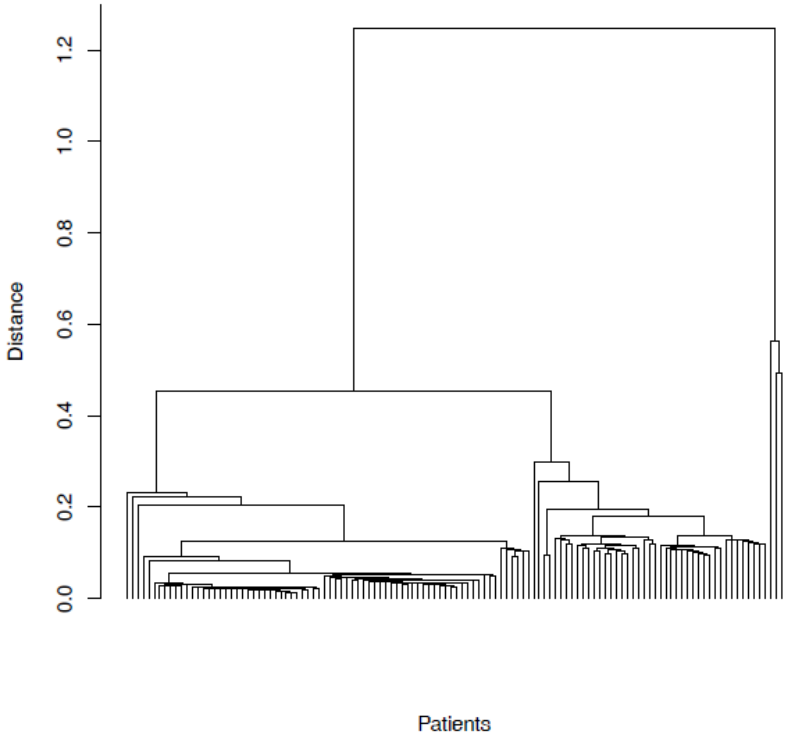
	Cluster 1	Cluster 2	p-value
N	117	3	
Sputum Eosinophilia†, %	8.91 (15.332)	38.28 (2.086)	0.008
Log Eosinophil count†, log %	0.34 (0.829)	1.58 (0.0236)	0.033
LABA(percentage=YES) ς	95.76%	50.00%	0.062
Exacerbations in last 12 months†,	3 (1-5)	0 (0-0)	0.048

† p-value derived from Mann-Whitney test, ς p-value derived from chi squared test

8.7.3 Factor 2 Eosinophilic Inflammation

Two clusters were detected on the second factor as can be seen in the cluster dendrogram, see graph 8.7.3.1. On inspection of the density distribution for factor 2, see graph 8.7.3.2, it is possible to see the multimodal nature of the distribution, mean dip statistic=0.02675 (p=0.10) and possibly 2 or 3 groups, so if using p=0.05 as a cut off factor 2 is multi modal. The clusters seen here were similar to that of factor one being significant for log eosinophil count suggesting a sub-group with large amount of eosinophilia and a group with normal eosinophilia see table 8.7.3.1

Graph 8.7.3.1 Cluster dendrogram for factor 2 (eosinophilic inflammation) using the probability of being in a cluster with another patient to separate patients.



Graph 8.7.3.2 Histogram of the density of latent factor 2 (eosinophilic inflammation)

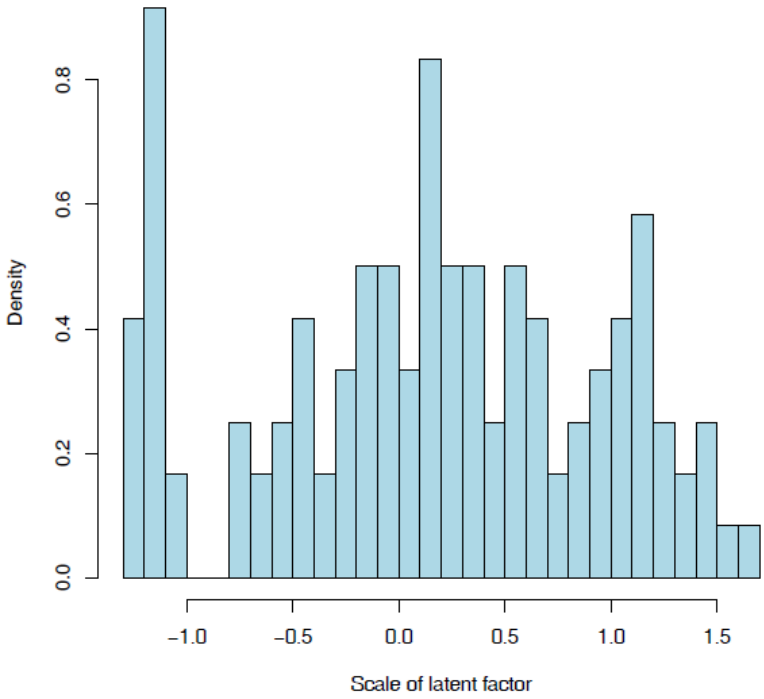


Table 8.7.3.1 The significant variables for the clusters found in factor 2 (eosinophilic inflammation)

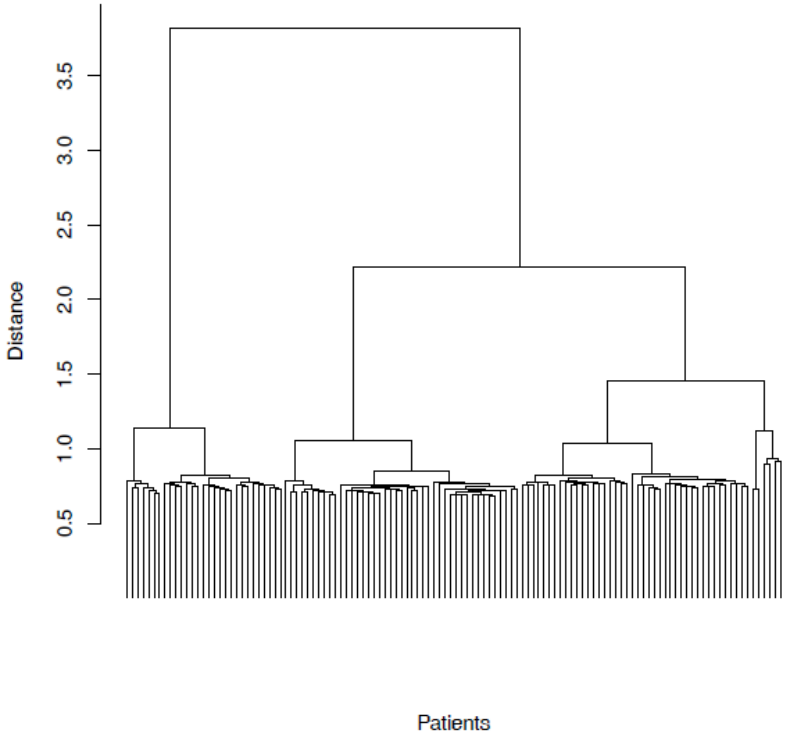
	Cluster 1	Cluster 2	p-value
N	117	3	
Sputum Eosinophilia†, %	8.93 (15.40)	27.43 (18.84)	0.043
Log Eosinophil count†, log %	0.337 (0.831)	1.308 (0.475)	0.039

† p-value derived from Mann-Whitney test

8.7.4 Factor 3: Air Flow Obstruction

The third factor had cluster dendrogram as in graph 8.7.4.1 this implied the data was separated into two groups again. Inspection of the density distribution of factor 3, see graph 8.7.4.2, suggests that the distribution is multimodal but this is confirmed by the mean as it is significant for this factors distribution, mean dip statistic = 0.02543 (p=0.05), but only just . The two clusters obtained are significant for FEV1 % predicted both pre and post, cell counts for neutrophils and JAC symptom score. The clusters are also significant for gender height and weight. Cluster 1 describes an air flow obstructed group with larger amounts of symptoms and a larger amount of neutrophilic inflammation who are are taller and heavier, and are predominantly male they also had a low FEV1/FVC ratio then group 2 which consisted of patients with good spirometry measurements less neutrophilic inflammation and less symptoms who were shorter and weighed less.

Graph 8.7.4.1 Cluster dendrogram for factor 3 (air flow obstruction) using the probability of being in a cluster with another patient to separate patients.



Graph 8.7.4.2 Histogram of the Density of latent factor 3 (Air flow Obstruction)

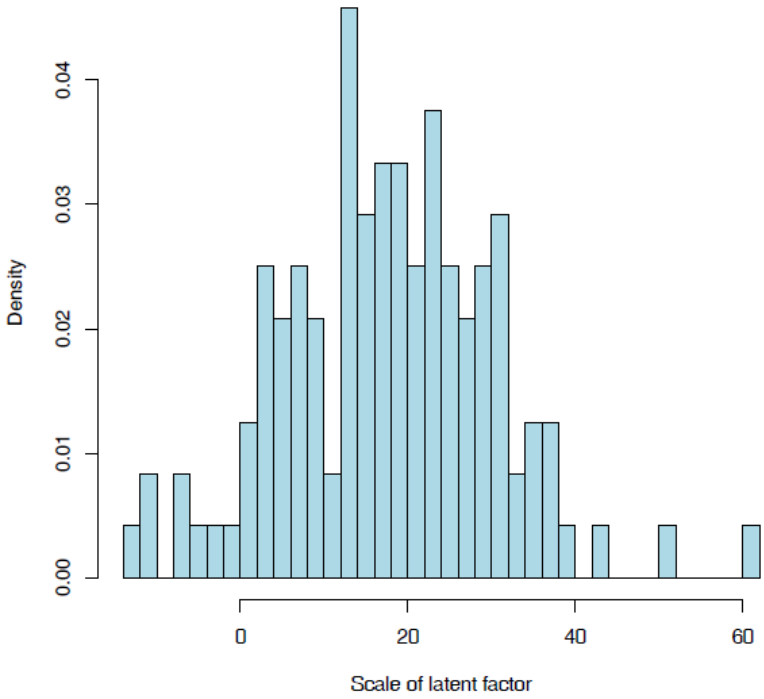


Table 8.7.4.1 the significant variables for the clusters found for factor 3 (air flow obstruction)

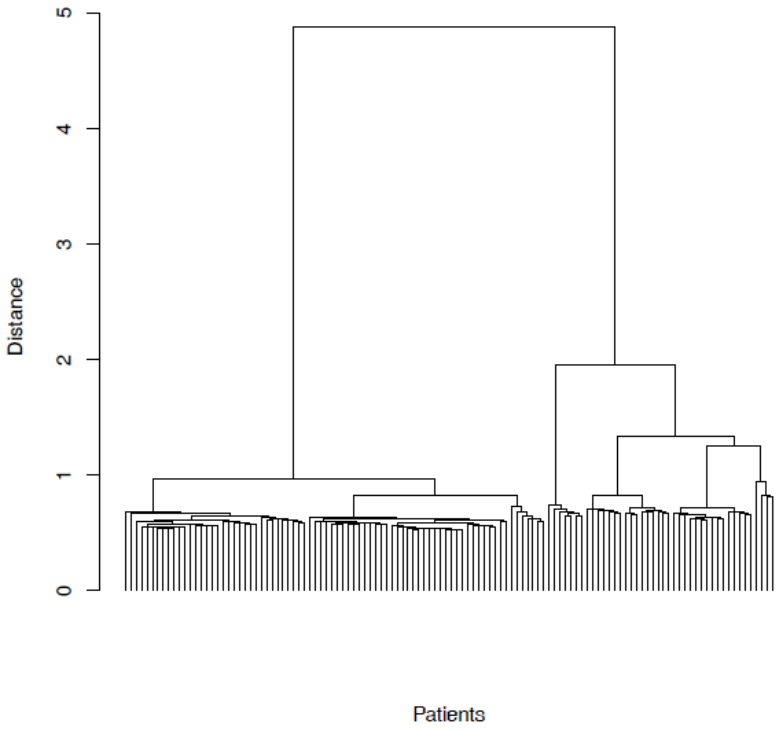
	Cluster 1	Cluster 2	p-value
N	91	29	
JACS†, mean score	2.84 (1.14)	2.07 (1.47)	0.005
Pre FEV1/ Predicted*, %	52.29 (16.64)	81.58 (16.92)	<0.001
Post FEV1/ Predcited*,%	57.48 (17.36)	84.94 (18.45)	<0.001
Pre FEV1/FVC*, %	0.55(0.11)	0.72(0.10)	<0.001
Post FEV1/FVC*, %	0.56(0.11)	0.74 (0.08)	<0.001
Pre FEV1*, Ls^{-1}	1.83(0.80)	2.31(0.72)	0.002
Post FEV1*,Ls^{-1}	1.99 (0.84)	2.45 (0.72)	0.005
Sp Neutrophilic*, %	67.11 (26.05)	55.33 (26.28)	0.037
Log Neutrophilic †, log %	1.77 (0.2544)	1.66 (0.3330)	0.036
Sp Macrophages†, log %	18.16 (19.08)	31.33(38.20)	<0.077
Gender (percentage male) ς	75.86%	27.47%	<0.001
Height (cm)*	173(8.65)	164(8.97)	<0.001
Weight (kg)*	91.91 (21.73)	82.52(20.51)	0.036

*p-value derived from one way ANOVA, ς p-value derived from chi squared test, † p-value derived from Mann-Whitney test

8.7.4 Factor 4 BMI

The fourth factor has 2 prominent clusters in its cluster dendrogram, see graph 8.7.4.1, this implied the data was separated into two groups, but inspection of the density distribution of factor 4,see graph 8.7.4.2, suggests that these two groups overlap and are not as clearly visible as seen for other groupings, and it is difficult to determine if the two groups make up a non-normal continuous distribution that can be approximated using two normally distributed mixtures or if the two mixtures make up two distinct sub-groups. The mean dip statistic= 0.02434 (p=0.05) implies a uni-modal distribution with only 33.53 % of iterations passing the dip test, confirming that the two mixtures are not distinct groups. In this dataset the two clusters split up into a younger mainly female obese group which have a lower IgE but higher FEV1/FVC ratio and an older non-obese group that had a higher IgE count who are more likely to be male and have a smaller FEV1/FVC ratio, see table 7, but it worth noting that the clusters seen here are not true clusters but come from an artificially split non-normal distribution.

Graph 8.7.4.1 Cluster dendrogram for factor 4 (BMI) using the probability of being in a cluster with another patient to separate patients.



Graph 8.7.4.2 Histogram of the density of latent factor 4 (BMI)

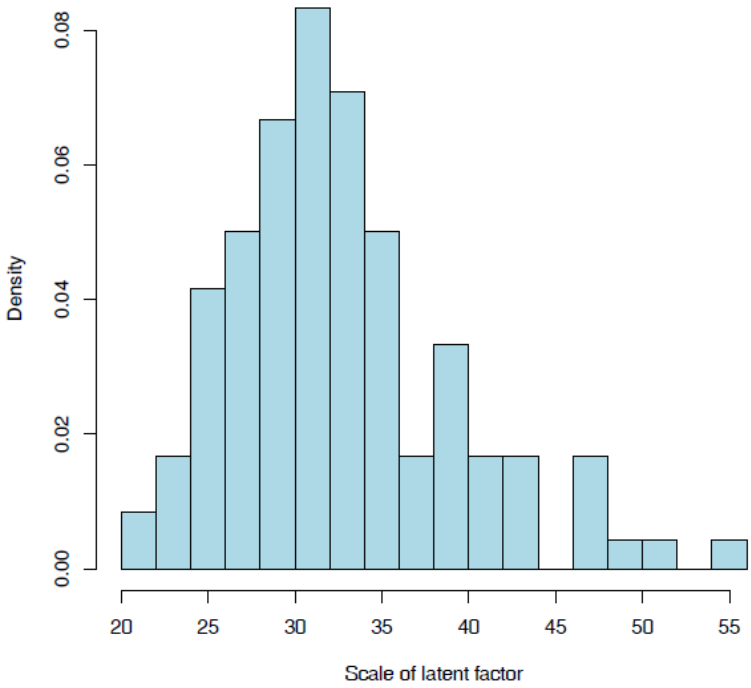


Table 8.7.4.1 The significant variables for the clusters found in factor 4 (BMI)

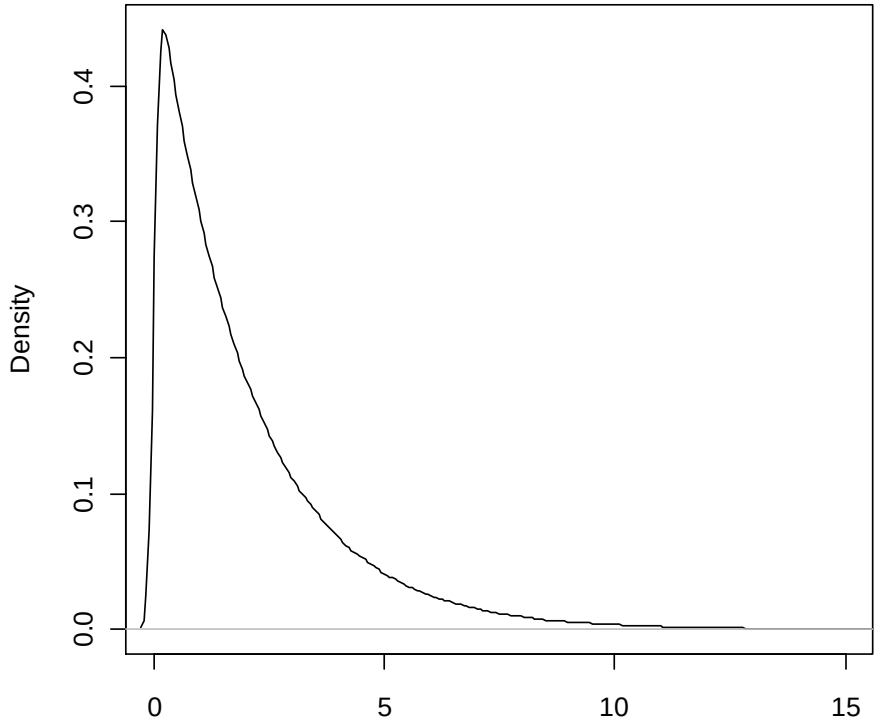
	Cluster 1	Cluster 2	p-value
N	42	78	
BMI*, kg/m^2	37.96 (6.08)	26.65 (3.62)	<0.001
Total IgE*, kU/l	270 (477)	340 (518)	0.022
Gender Percentage male $^{\varsigma}$	26.19%	46.15%	0.030
Weight (kg)*	103.54(19.92)	74.69 (13.51)	<0.001
Age (yrs)*	47.56 (10.96)	52.31 (13.26)	0.05
Pre FEV1/FVC*, %	0.71 (0.11)	0.66 (0.13)	0.029
Post FEV1/FVC*, %	81.26 (19.68)	76.71 (22.56)	0.023

*p-value derived from one way ANOVA, $^{\varsigma}$p-value derived from chi squared test

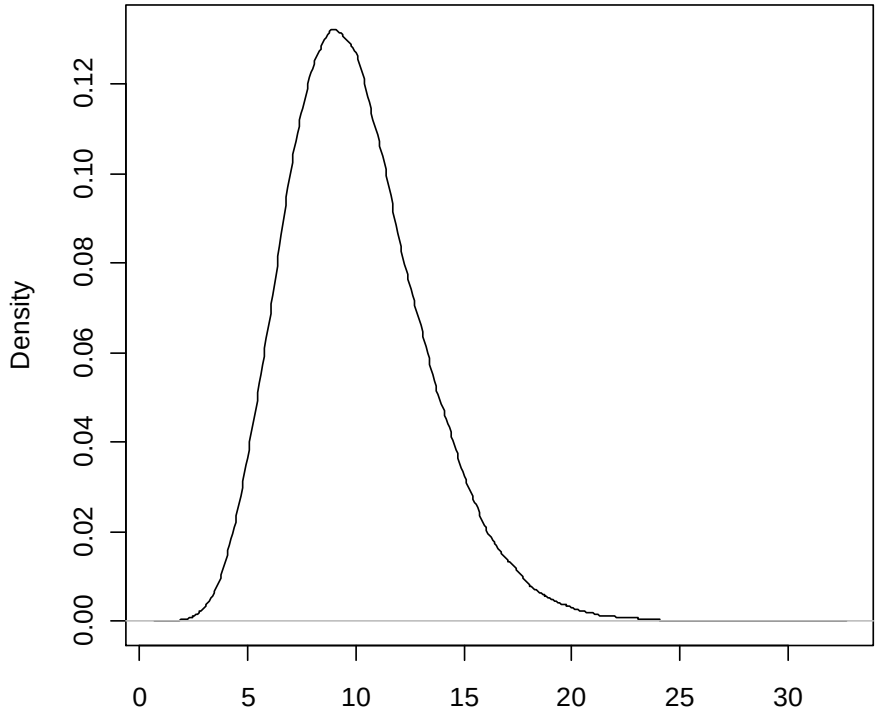
8.8 Specificity analysis on alpha parameter

The original analysis was carried out using the gamma distribution as a prior for the parameter α with shape=1 and scale =2, see graph 8.7.4.1, as in the simulations. The α parameter is usually dictated by the data when using large datasets. In order to check the prior on the α parameters effect two other priors were used for the analysis keeping all other parameters constant. The two new priors used were giving the α parameter a strict single value α=1 which is common in Dirichlet Process and Dirichlet process mixture analysis. The other prior used in the specificity analysis was a gamma distribution that favoured production of more mixtures this had a gamma distribution with shape=10, scale=1, see graph 8.7.4.2.

Graph 8.7.4.1: Original prior distribution of alpha parameter used in previous analysis and simulation. Gamma distribution with shape=1,scale=2.



Graph 8.7.4.2: New prior distribution for alpha parameter used in sensitivity analysis. Gamma distribution with shape=10 and scale =1.



8.8.1 Results from sensitivity analysis

The percentage of parameters that had converged was 18.61% for the model with the alpha parameter equalling 1 and 70.11% using the model with the alpha parameter having a gamma prior with shape=10 and scale=1 suggesting that the alpha prior was possible to strict for the smaller dataset.

The same number of clusters, 2 was returned for each factor using each prior. Similar densities were found for all factors across priors, see Appendix 7. Cluster membership was similar for each factor over the priors, see Appendix 7 for cluster dendrograms under prior alpha=1 and for the new gamma prior shape=10 scale=1.

Cluster membership correlation was carried out using the Kappa test for ordinal correlation to quantify the amount of reliability in cluster membership, see table

8.7.5.2 The original prior and new prior with gamma shape=1 and scale=10 had significant similar cluster memberships for factors 1 to 3, but for factor 4 (BMI) however they differed, but still produced an obese cluster and a non obese cluster. The alpha=1 prior seems to show the most difference then the other two priors, it is similar to the mixtures for factors 2 and 3 but was different for factors 1 and 4. This suggests that the priors are possibly having an impact on the clustering due to the smaller amount of data.

Table 8.7.5.1 Effect of prior specification on alpha for number of clusters found.

Membership	Factor 1	Factor 2	Factor 3	Factor 4
Original	2	2	2	2
Alpha	2	2	2	2
New	2	2	2	2

Table 8.7.5.2 Best Kappa correlation of cluster membership for each comparison of priors

Membership	Factor 1	Factor 2	Factor 3	Factor 4
Original VS Alpha	-0.026 (p=0.744)	1.00 (p<0.001)	0.809 (p<0.001)	0.057 (p=0.479)
Original VS New	1.00 (p<0.001)	1.00 (p<0.001)	0.809 (p<0.001)	0.028 (p=0.522)
New VS Alpha	-0.026 (p=0.744)	1.00 (p<0.001)	1.00 (p<0.001)	0.310 (p<0.001)

8.8 Discussion

Clusters were found on all factors which were significant for asthma outcome variables however when using the mean dip test to verify if the clusters made up a multi-modal distribution or not most of the clusters proved to be non significant. The only multi-modal distribution accordingly to the mean dip statistic was that for the eosinophilic factor producing two group's one with a large amount of eosinophils and one with a normal amount of eosinophils. Although the clusters found on factors 1, 3 and 4 were significant for some asthma outcomes the dip statistic was to low resulting in the distribution passing the test for uni-modality. The obese sub-group has been found yet again to be part of a bigger non-normal distribution as has airflow obstruction and lung volume.

As the data only have 120 patients in it, it is possible that the dataset is under powered this would explain the multimodality being found and the different results for using the priors too. If more data were added or a new larger dataset used this could clear the question mark over the clusters nature.

8.9 Closing statement

Although we have seen that it is possible to find mixtures over factors and that these mixtures show significant differences in a number of asthma related variables. The mixtures seem to reflect a bigger non normal distribution rather than specific mixtures which come from different populations. This is not usually tested during cluster analysis and is an important step to consider when interpreting mixtures and clusters. As datasets may be too small to determine if clusters are overlapping mixtures making non normal distributions or genuine clusters. In the case of the Brompton dataset its seems that it is a little too small to test for clusters, although an eosinophilic/ non

eosinophilic split was found. We now look at a different use for Dirichlet Process Normal Mixture Latent Variable Models for non conjugate data to determine sub-groups or clusters in a clinical trial

Chapter 9. Analysis of Haldar Severe Asthma Dataset

9.1 Chapter outline

The Haldar dataset is now analysed using the Dirichlet Process Normal Mixture Latent Variable Model DPNMLVM and the truncated Dirichlet Process Normal mixture Latent variable Model TrDPNMLVM. In order to compare it with the previous k-means cluster analysis. Four factors were found that are similar to the factors that were found in the previous analysis. Clusters were found on the factors with multi modal mixtures being found on the atopic factor. Suggesting an atopic/non atopic split that was also previously found in the k-means cluster analysis. The splitting of the obese factor however resulted in two overlapping mixtures that make up a bigger non-normal distribution and do not represent separate clusters. The results were consistent over all three different priors for α. Suggesting that the sample size was adequate to detect mixtures that were not due to sampling. Results were similar for both the full and truncated model but some factors did differ between models as they picked up on different patterns of variation within the cohort.

9.2 Introduction

We now look at using the Dirichlet Process Normal Mixture Latent Variable Model (DPNMLVM) and the TrDPNMLVM to analyse the Haldar severe asthma dataset. Using the DPNMLVM/TrDPNMLVM it is possible to determine the shape of the distribution of each independent factor and the nature of the mixtures on each of the factors that contribute to the patient variation. It is also possible to determine a clustering partition to examine the existence of sub-groups on the severe asthma factors and to

annotate these depending on which variables these partition. The DPNMLVM methodology used in the analysis is the same as outlined and tested in the previous two simulation chapters, chapter 6 and 7 for continuous data variables that can be correlated or uncorrelated. The TrDPNMLVM was used using a similar method but limiting the number of mixtures to a finite amount, 15. Although setting boundaries on the number of distributions is a limitation, the maximum number of groups allowed can be set to be much higher than is thought likely to occur and 15 was found to be sufficient. As long as the number of mixtures fixed a prior N is larger than the actual number of mixtures needed so that the Dirichlet process mixture converges to, then the algorithm works in a very similar way then the full Dirichlet process. The truncated/approximate Dirichlet process mixture can be used to compute a very good approximation to the full Dirichlet process mixture which is faster than marginal methods (Ishwaran and Zarepour 2000) although they should give similar results they are both different algorithms and for this reason we compare them for the Haldar Analysis.

9.3 Haldar dataset

The first dataset we will look at is the Pranab Haldar severe Asthma dataset taken from (Haldar, 2007) we used this to demonstrate the DPNMLVM as the dataset has been used before on a previous k-means cluster analysis that had produced four clusters and it would be interesting to see if the methods produced similar clusters and results.

The dataset used contains demographics and asthma biomarkers on 187 subjects. All subjects were attending the difficult asthma clinic at the Glenfield Hospital, Leicester and have a diagnosis of refractory asthma in accordance with the American Thoracic Society (ATS) criteria at the time of its publication although possible on reduced

amount of inhaled corticosteroids, to see the dataset statistics see chapter 2 for the dataset descriptions.

The Haldar k-means cluster analysis used the continuous variables; Modified JACSs symptom score, BMI, logged sputum eosinophil counts and age of onset and the binary variables; atopic (yes/no) and gender male/ female. The k-means clustering algorithm is built to partition continuous variables into clusters but the nature of these clusters is unknown as they can only be described using multivariate statistics and not scaled diagrams due to the large number of dimensions. Uni-variate distributions for each variable can be used to display the clusters but do not reflect the true nature of the multivariate clusters.

The partition created during the k-means cluster analysis could depend on the misused binary variables that are not commonly used to determine clusters in a continuous clustering algorithm such as k-means clustering (Everitt 2001). In order to replicate the analysis without the use of binary variables I used the 9 continuous variables listed below, most are the same as seen in the Haldar analysis, but to replace the binary variable describing atopy, skin prick test data was used to mimic the atopic status as the binary atopic variable is a dichotomised version of this data. The binary variable gender was taken out of the full DPNMLM analysis as the derived methodology for the full DPNMLM does not work with binary variables and also gender although affecting a patient's asthma is not a direct outcome of asthma. However gender as a variable was included as an extra step in the TrDPNLVM to see if this variable correlated to any of the factors. As this truncated methodology allowed binary variable to be included.

9.4 Variables

The 9 variable that were used in the DPNMLVM analysis are as follows; 4 skin prick tests for dog, cat, and grass pollen and house dust mites respectively, which measure the reaction to each of these substances to determine allergies for these substances. Modified JACS Symptom Scores which is a symptom questionnaire devised for asthma symptoms with the FEV1 part of the score removed, body mass index (BMI), log of sputum eosinophil count, log of sputum neutrophil count, both eosinophils and neutrophils are inflammation cells found in severe asthma sputum samples and age of onset of symptoms of asthma.

9.5 Classical factor analysis

Factor analysis was carried out on the variables under the usual assumption of normally distributed latent/factor variables to determine the number of latent variables and which variables would be factor/latent variable anchors in the DPNMLVM as in the previous simulation chapter, chapter 7. The normally distributed factor analysis was carried out using SPSS. The number of factors was chosen by selecting factors with eigenvalues greater than 1 to 2 decimal places known as the Kaiser criteria. Four factors were found to be sufficient which explained around 72% of the variance seen in the dataset. See table 9.5.1. The four factor anchors were chosen by selecting the variables that had the highest factor loading on a factor, i.e. the variable that had the highest correlation with a factor. See table 9.5.2.

Table 9.5.1: The results of the standard factor analysis of the 9 variables from the Haldar severe asthma dataset. A four factor solution satisfies the Kaiser criteria and allows for 72% of the variance.

No of Factors	Eigenvalues	Cumulative % of Variance
1	3.176	35.291
2	1.297	46.699
3	1.014	60.961
4	0.995	72.012
5	0.881	81.798
6	0.656	89.093
7	0.481	94.438
8	0.345	98.270
9	0.156	100.000

Table 9.5.3 shows the factor loading for the four normally distributed factor model. The highest factor loading for each variable has been highlighted and used for annotation of the factors. The highest loading variable for each factor was used as a factor anchor for that factor in the DPNMLVM.

Variables	1	2	3	4
BMI kg/m^2	-.098	.428	.450	.590
Age of Onset of Symptoms, years	-.577	-.163	-.415	.267
Log eosinophil cell count log %	-.039	.622	-.526	.332
Modified JACS symptom score	-.217	.425	.652	-.357
Log neutrophil cell count log %	-.014	-.462	.456	.581
Cat skin prick test, mm	.881	.019	.060	-.034
Dog skin prick test, mm	.876	.065	-.028	.099
House Dust Mite skin prick test, mm	.721	-.019	-.053	.070

Grass Pollen Skin Prick Test, mm	.850	-.002	-.068	.056

The factors can be described by the variables that they correlate to. Thus Factor 1 describes an atopy component to the dataset. Factor 2 describes an eosinophilic inflammation factor. Factor 3 describes the symptomatic aspect to asthma and factor 4 describes the role of BMI in the dataset.

The anchors for each factor to be used in the DPNMLM are the highest factor loading variable for each factor these are cat skin prick test, log eosinophil cell count, Modified JACS symptom score and BMI. This is the rationale for using four factors with the four factor anchors in the DPNMLM.

9.6 Dirichlet Process Normal Mixture Latent Variable Model

For variables j for j=1,2,…9

For subjects i for i=1,2...120

For latent variable l for 1,2...4

$$Y_{ij} \sim N\left(\mu_{ij}, {\sigma_j}^2\right) \qquad \textit{Equation 202}$$

$$\mu_{ij} = \beta_{0j} + \beta_{lj}.Z_{li} \qquad \textit{Equation 203}$$

$$Z_{li} \sim D_l(\alpha_l, G_{o_l}) \qquad \textit{Equation 204}$$

$$G_{lo} \sim N(\theta_{li}, V_{li})$$ *Equation 205*

Where Y_{ij} represents the i individual of the j normally distributed variables μ_{ij} represent the mean of the Y_{ij}, ${\sigma_j}^2$ is the variance of the Y_{ij} variable, β_{0j}, β_{lj} parameters of the regression of μ_{ij} on latent variables Z_{li} , $D_l(\alpha_l, G_{lo})$is the Dirichlet process mixture over latent variable l with precision parameter α_l and centring distribution G_{lo}, where G_{lo} is normally distributed with mean θ_{li} and variance V_{li}. β_{01}, β_{02}, β_{03}, β_{04} all equal 0 and β_{11}, β_{22}, β_{33}, β_{44} all equal 1 as these are the parameters associated with the factor anchors and kept constant for identity purposes, where

Y_{i1} = Cat skin prick test

Y_{i2} =Log eosinophil cell count

Y_{i3} = Modified JACS score

Y_{i4} =BMI

Y_{i5} =Dog skin prick test

Y_{i6} =House dust mite skin prick test

Y_{i7} =Grass pollen skine prick test

Y_{i8} =Age of onset of symptoms

Y_{i9} = Log neutrophil cell count

9.7 Priors and Convergence

Priors were kept the same as in the previous simulation chapter, with a sensitivity analysis carried out later on the alpha parameter of the model. The model parameters were tested for convergence using the Heidelberg test after 425,000 iteration burn in and for a further 75,000 iterations after. The percentage of parameters that passed this test was 64.39%. The parameters that did not pass this test were checked for convergence.

9.8 Results for Dirichlet process normal mixture model

9.8.1 Variable analysis on factors

The correlation for each variable against the four factors of the DPNMLVM can be seen in table 9.8.1 this is similar to the factor loading table from table 2 for the normally distributed factor analysis, suggesting that the factors are the same and the analysis is similar. Factor 1 corresponded to atopy, factor 2 to eosinophilic inflammation, factor 3 to modified JACS symptom score and factor 4 to BMI. With the factors established we now turn our attention to the clusters in each factor.

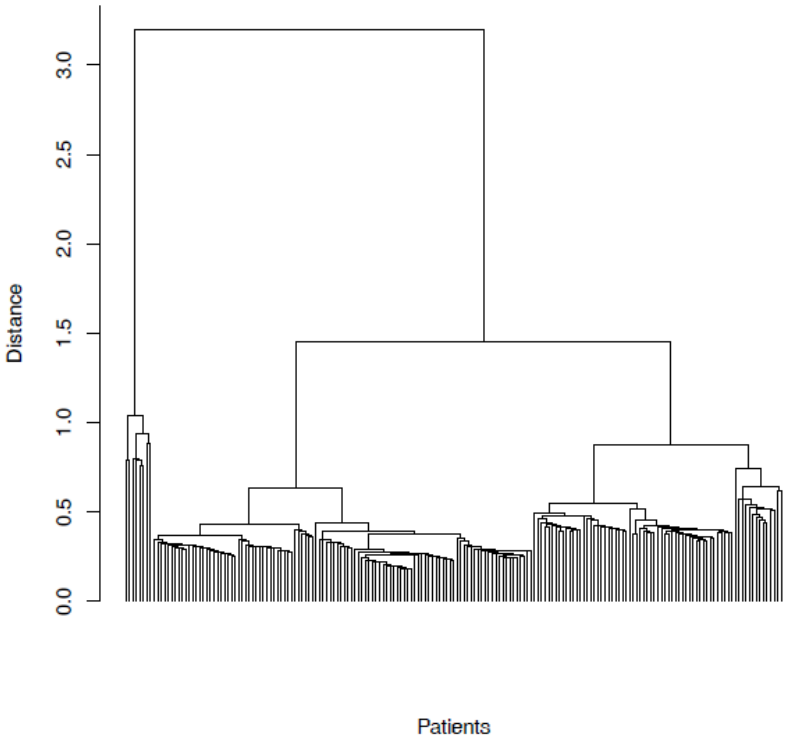
Table 9.8.1: The correlation of the variables with the factors in the Dirichlet Process Normal Mixture Latent variable model, the highest factor loading for each variable is highlighted

Variables	1	2	3	4
BMI, kg/m^2	-0.077	0.107	0.083	0.987
Age of Onset of Symptoms, years	-0.466	0.101	-0.151	-0.036
Log eosinophil cell count, log %	-0.011	0.991	0.006	-0.104
Modified JACS symptom score	-0.081	-0.010	0.994	0.019
Log neutrophil cell count, log %	0.001	-0.093	0.032	0.113
Cat skin prick test, mm	0.963	-0.043	0.002	-0.008
Dog skin prick test, mm	0.916	0.081	-0.109	0.024
House Dust Mite skin prick test, mm	0.573	-0.038	-0.212	0.076
Grass Pollen Skin Prick Test, mm	0.798	0.044	-0.133	-0.032

9.8.2 Factor 1: Atopy

The first factor had cluster dendrogram as in graph 9.8.2.1 suggesting the factor is separated into two clusters. Inspection of the density distribution for factor 1, See graph 9.8.2.2, suggests that the distribution is multi modal indicating that the clusters found are real sub-groups. This is further backed up by the distribution passing the dip test of multi-modality as we retain the null hypothesis of the distribution being multimodal, dip statistic = 0.02395. (p=0.5). The clusters found are significantly different for means of age, age of onset cat skin prick test, dog skin prick test, grass pollen skin prick test and house dust mite skin prick test, see table 9.8.2.1. The data in factor 1 is split into cluster 1 which describes an older group with a later onset of disease whose patients are less atopic then cluster 2 which has a lower mean age of onset and contains younger patients who are highly atopic.

Graph 9.8.2.1: Cluster dendrogram for factor 1 using the probability of being in a cluster with another patient to separate patients.



Graph 9.8.2.2 Histogram of the density of latent factor 1 (atopy)

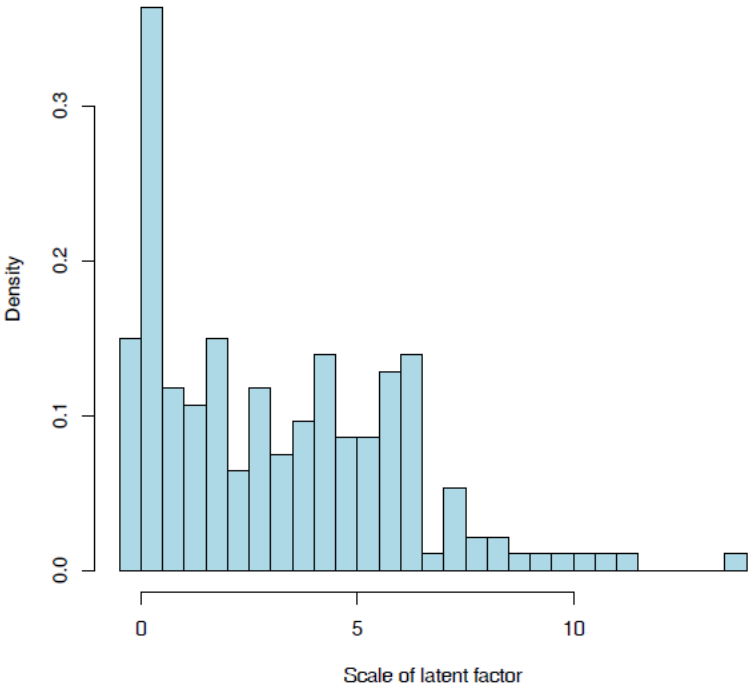


Table 9.8.2.1: The significant variables for the clusters found in factor 1 (atopy)

	Cluster 1	Cluster 2	P-Value
N	179	8	
Age *, years	43.97 (15.9)	31.50 (11.3)	0.030
Age of Onset*, years	20.88	6.50	0.013
Cat†, mm	2.75	8.63	0.001
Dog†, mm	2.84	9.63	<0.001
Grass†, mm	3.22	8.13	0.010
Dust mite†, mm	3.69	6.00	0.068

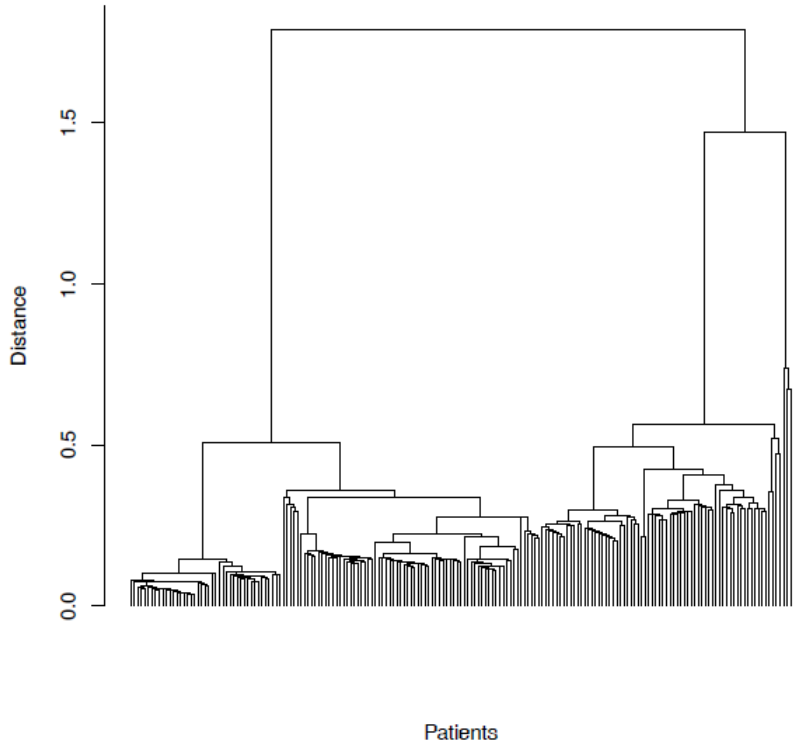
*p-value derived from t-test, † p-value derived from Mann-Whitney test

9.8.3 Factor 2: Eosinophilic inflammation

Three clusters were detected in the second factor as can be seen in the cluster dendrogram, see graph 9.8.3.1 On inspection of the density distribution for factor 2, see graph 9.8.3.2 however suggests that these three groups are not easily visible, and that the three groups could make up a non-normal continuous distribution that can be approximated using three normal distributions, this is further backed up by the dip statistic which although being non-significant for this factor density distribution had a very low dip statistic, 0.01947 (p=0.10), so if using p=0.05 as a cut off factor 2 is multi modal but not by a large degree. The only variable that was significantly associated with the eosinophilic factor other than the factor itself was gender, see table 9.8.3.1. The clustering suggested a cluster that consisted of females that have lower eosinophilic inflammation then the other two clusters that have similar eosinophil

levels. This suggests that the two groups with similar eosinophilia probably make up a non-normal distribution with the possible third group being a genuine sub-group consisting of females with lower eosinophil counts.

Graph 9.8.3.1: Cluster dendrogram for factor 2 (eosinophilic inflammation) using the probability of being in a cluster with another patient to separate patients.



Graph 9.8.3.2: Histogram of the density of latent factor 2 (eosinophilic inflammation)

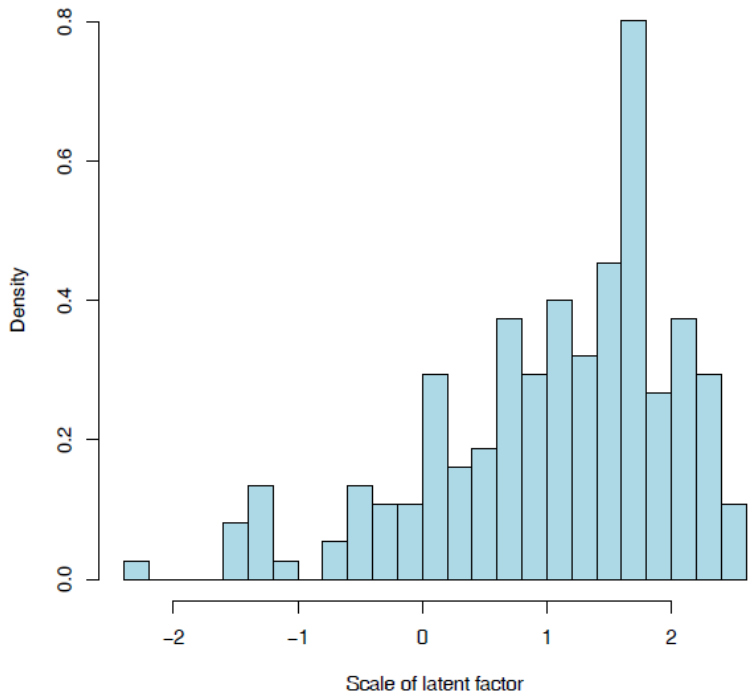


Table 9.8.3.1: The significant variables for the clusters found in factor 2 (eosinophilic inflammation)

	Cluster 1	Cluster 2	Cluster 3	p-value
N	121	63	3	
Sex ᶜ (%=male)	28.45%	45.59%	0%	0.018
Log eosinophils*, log %	0.47 (1.02)	0.45 (0.95)	0.34 (1.34)	0.967

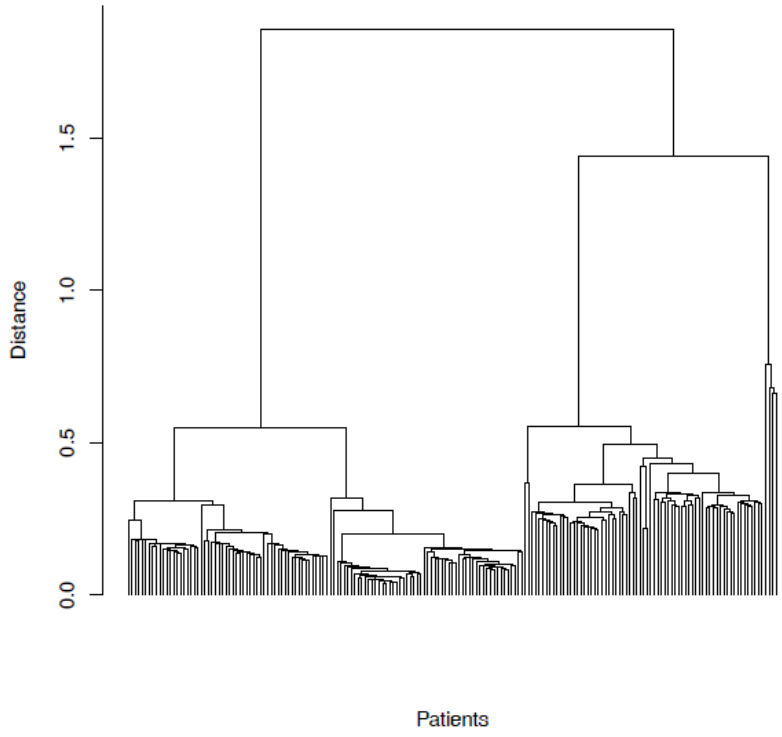
*p-value derived from one way ANOVA, ᶜ p-value derived from chi squared test

9.8.4 Factor 3: Symptoms

The third factor had cluster dendrogram as in graph 9.8.4.1 this implied the data was separated into three groups again. Inspection of the density distribution of factor 3,

see graph, 9.8.4.2 suggests that the distribution is multimodal and this is further backed up by the dip test being non-significant for this factors distribution, dip statistic = 0.02109 (p=0.5). Again as with the eosinophilic inflammation cluster the three clusters obtained are only significant with gender with the third cluster describing a female only cluster with a larger number of symptoms. Suggesting that gender is again correlated to the symptom factor as well as the factor correlated with inflammation.

Graph 9.8.4.1: Cluster dendrogram for factor 3 (Symptoms) using the probability of being in a cluster with another patient to separate patients.



Graph 9.8.4.2: Histogram of the Density of latent factor 3 (symptoms)

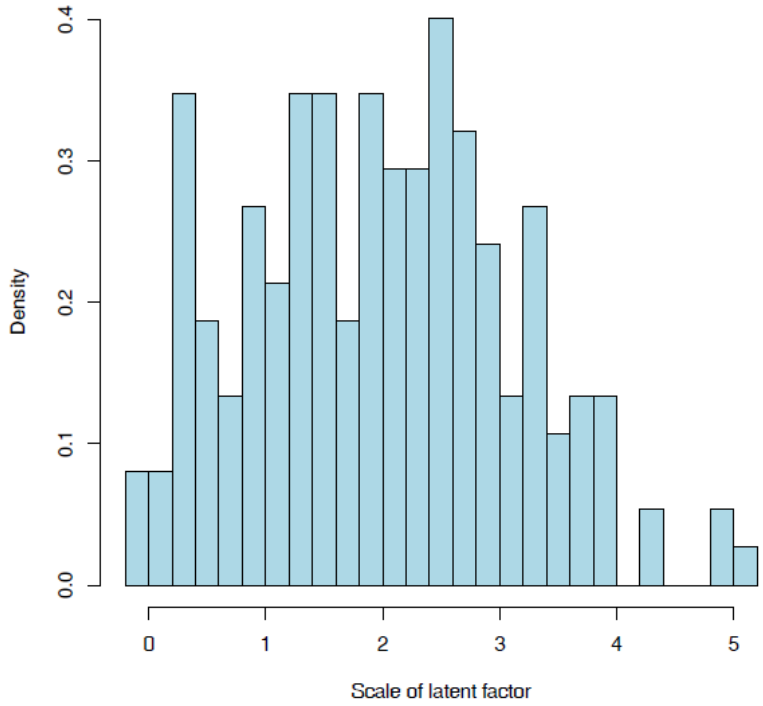


Table 9.8.4.1: The significant variables for the clusters found in factor 3 (Symptoms)

	Cluster 1	Cluster 2	Cluster 3	p-value
N	120	63	4	
Sex ς (% male)	28.95%	44.93%	0%	0.017
JACS† mean score	2.00 (1.13)	2.04 (1.20)	2.48 (1.39)	0.706

ς p-value derived from chi squared test, † p-value derived from Kruskal-Wallis test

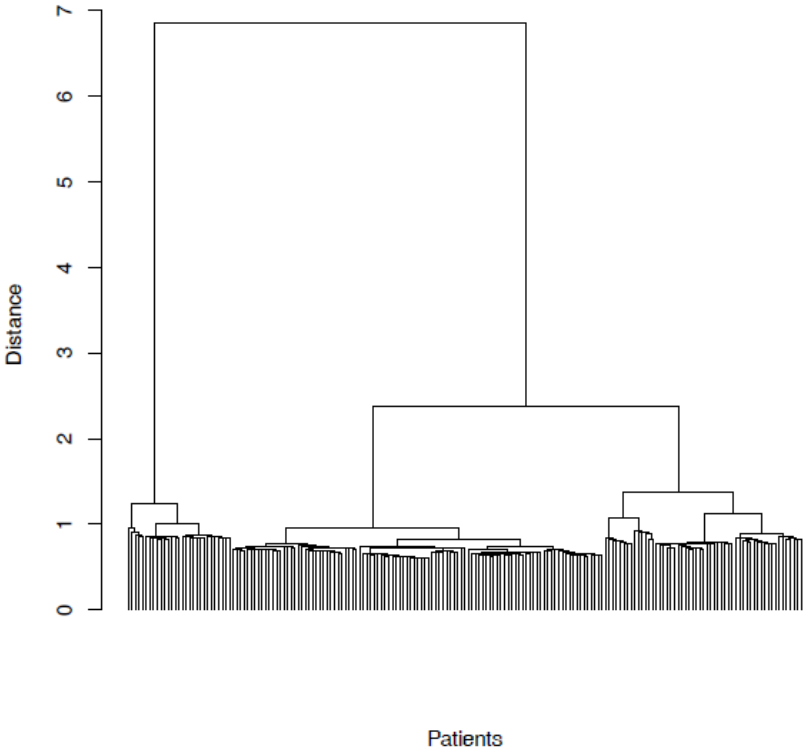
9.8.5 Factor 4: BMI

The fourth factor had 2 prominent clusters in its cluster dendrogram, see graph

9.8.5.1, this implied the data was separated into two groups, but inspection of the

density distribution of factor 4, see graph 9.8.5.2, suggests that these two groups are not easily visible and that the two groups make up a non-normal continuous distribution that can be approximated using two normally distributed mixtures this is further backed up by the dip statistic being non-significant for the factor density distribution, dip statistic = 0.01836, (p=0.05) confirming that the data is not multi modal. The two clusters split up into an obese group which were mainly female and a non-obese group that were more evenly distributed and used their bronchodilators less often, see table 9.8.5.1.

Graph 9.8.5.1: Cluster dendrogram for factor 4 (BMI) using the probability of being in a cluster with another patient to separate patients.



Graph 9.8.5.1: Histogram of the density of latent factor 3 (BMI)

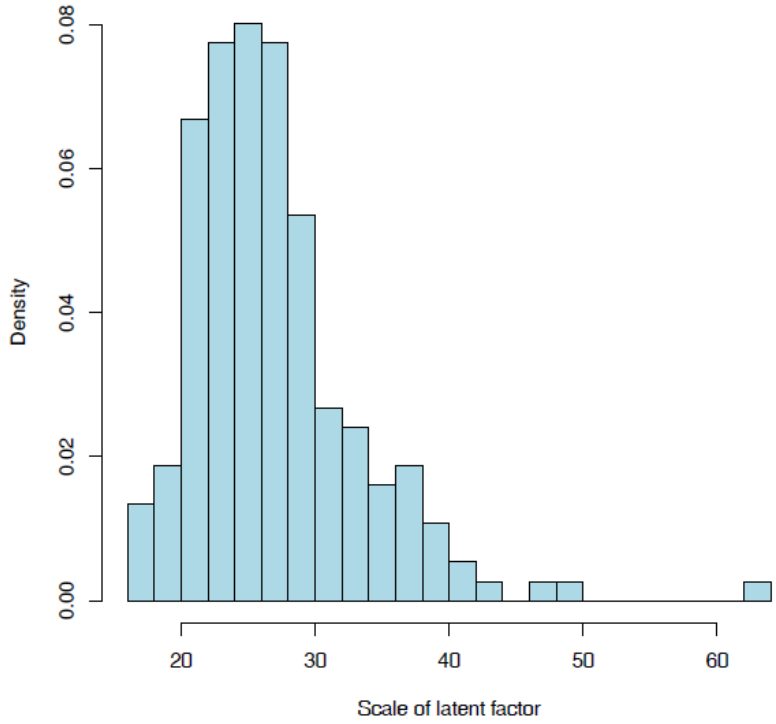


Table 9.8.5.1: The significant variables for the clusters found in factor 4 (BMI)

	Cluster 1	Cluster 2	p-value
	157	30	
Sex[ς] (% female)	38.61%	10.34%	0.001
BMI* kg/m^2	26.40 (3.84)	40.06 (5.96)	<0.001
Bronchodilator use*	5.94 (5.66)	10.07 (11.57)	0.018

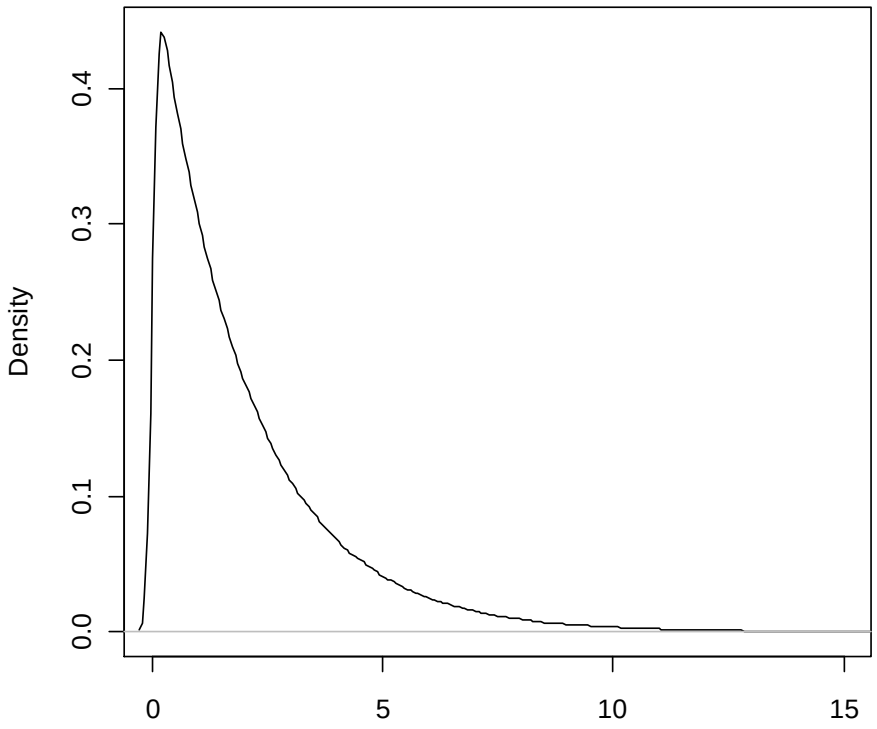
[ς] p-value derived from chi squared test, *p-value derived from t-test

9.9 Specificity analysis on α parameter

The original analysis was carried out giving the α parameter a gamma distribution as a prior with shape =1 and scale =2, see graph 8.9.1, as in the simulations. The alpha parameter is usually dictated by the data when using large datasets, but the data we have be insufficient to overcome the prior information. In order to check the prior on

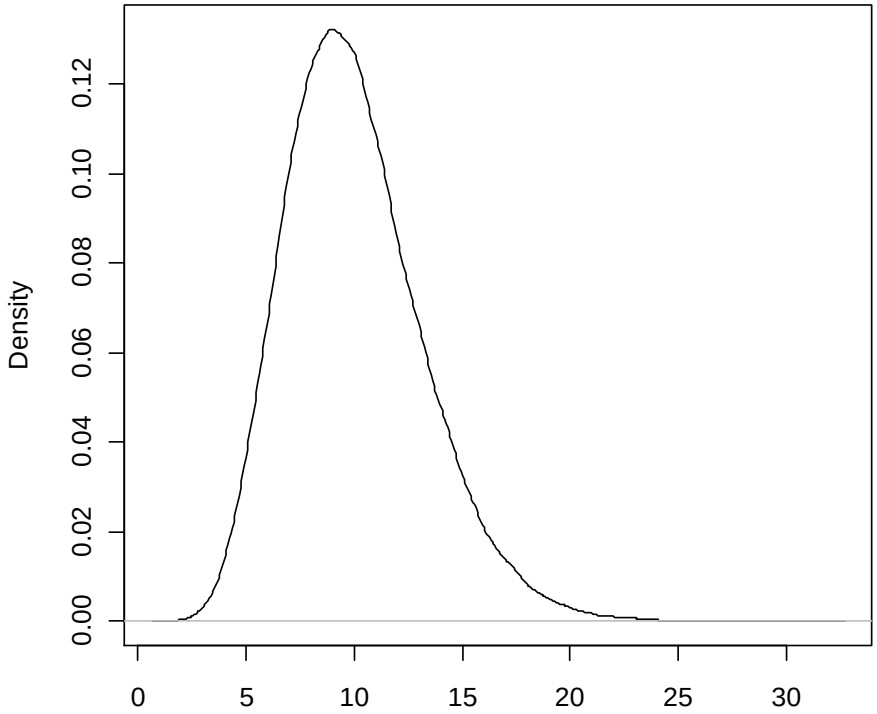
the α parameters effect two other priors were used for the analysis keeping all other parameters constant. The two new priors used for the α parameter were a strict single value α=1 which is common in Dirichlet process and Dirichlet process mixture analysis. The other prior used was a gamma distribution that favoured production of more mixtures this had a gamma distribution with shape=10, scale=1, see graph 9.9.2.

Graph 9.9.1: Original prior distribution of alpha parameter used in previous analysis and simulation. Gamma distribution with shape=1, scale=2.



Graph 8.9.2: New prior distribution for alpha parameter used in sensitivity analysis.

Gamma distribution with shape=10 and scale =1



9.10 Results from sensitivity analysis

The percentage of parameters that had converged was 66.47% for the model with the α parameter equalling 1 and 72.85% using the model with the α parameter having a gamma prior with shape=10 and scale=1.

The same number of clusters was returned for each factor apart from factor 2 using α=1 prior, see table 9.10.1. In this case the distances for cutting the dendrogram at 2 or 3 were very close with the distance for two clusters being slightly larger. Similar densities were found for all factors across priors, see graphs in appendix 6 for densities using the alpha=1 prior for factors 1, 2, 3 and 4 and for the new gamma prior with shape=10 and scale=1 for factors 1, 2, 3 and 4. Cluster membership was similar for

each factor, see appendix6 for cluster dendrograms under prior alpha=1 and cluster dendrograms for factors 1, 2 , 3 and 4 for the new gamma prior shape=10 scale=1.

Cluster membership correlation was carried out using the Kappa test for nominal correlation to quantify the amount of reliability in cluster membership. All factors with the same number of clusters were tested and all cluster memberships were significantly correlated to a high degree ($p<0.001$ in all cases with the same number of clusters) see table 9.10.2

Table 9.10.1: Effect of prior specification on alpha for number of clusters found.

Prior	Factor 1	Factor 2	Factor 3	Factor 4
Gamma(shape=1,scale=2)	2	3	3	2
Alpha=1	2	2	3	2
Gamma(shape=10,scale=1)	2	3	3	2

Table 9.10.2: Best Kappa correlation of cluster membership for each comparison of priors

Prior	Factor 1	Factor 2	Factor 3	Factor 4
Original VS Alpha	0.789 ($p<0.001$)	NA	0.945 ($p<0.001$)	0.657 ($p<0.001$)
Original VS New	0.938 ($p<0.001$)	0.989 ($p<0.001$)	0.967 ($p<0.001$)	0.837 ($p<0.001$)
New VS Alpha	0.748 ($p<0.001$)	NA	0.978 ($p<0.001$)	0.810 ($p<0.001$)

9.11 Truncated Dirichlet Process Normal Mixture Latent Variable Model

For variables j for j=1,2,…9

For subjects i for i=1,2...120

For latent variable l for 1,2...4

$$Y_{ij} \sim N(\mu_{ij}, {\sigma_j}^2) \qquad \textit{Equation 206}$$

$$\mu_{ij} = \beta_{0j} + \beta_{lj}.Z_{li} \qquad \textit{Equation 207}$$

$$Z_{li} \sim TrD_l(\alpha_l, G_{o_l}) \qquad \textit{Equation 208}$$

$$G_{lo} \sim N(\theta_{li}, V_{li}) \qquad \textit{Equation 209}$$

Where Y_{ij} represents the i individual of the j normally distributed variables μ_{ij} represent the mean of the Y_{ij}, ${\sigma_j}^2$ is the variance of the Y_{ij} variable, β_{0j}, β_{lj} parameters of the regression of μ_{ij} on latent variables Z_{li} ,$Tr\ D_l(\alpha_l, G_{lo})$is the truncated Dirichlet process mixture over latent variable *l* with precision parameter α_l and centring distribution G_{lo}, where G_{lo} is normally distributed with mean θ_{li} and variance V_{li}. $\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}$ all equal 0 and $\beta_{11}, \beta_{22}, \beta_{33}, \beta_{44}$ all equal 1 as these are

the parameters associated with the factor anchors and kept constant for identity purposes, where

Y_{i1} = Cat skin prick test

Y_{i2} =Log eosinophil cell count

Y_{i3} = Modified JACS score

Y_{i4} =BMI

Y_{i5} =Dog skin prick test

Y_{i6} =House dust mite skin prick test

Y_{i7} =Grass pollen skine prick test

Y_{i8} =Age of onset of symptoms

Y_{i9} = Log neutrophil cell count

9.12 Priors and Convergence

Priors for β_{lj} ${\sigma_j}^2$ θ_{li} and V_{li} were kept the same as in the previous analysis with the full Dirichlet process, The alpha prior was kept as a uniform distribution from 0.3 to 7 for identification purposes as suggested in (Ohlssen, Sharples et al. 2007). The maximum number of sub- groups was kept at 15 for each treatment group to obtain a fair approximation of the full Dirichlet process mixture. Previous models with a higher number of subgroups were used and were found to contain many empty groups.

The model parameters were tested for convergence using the Heidelberg test after 5,000 iteration burn in and for a further 95,000 iterations after with a thinning value of

5 , meaning every fifth iteration would be saved. The percentage of parameters that passed this test was 97.2%. The parameters that did not pass this test were checked for convergence.

9.13 Results for truncated Dirichlet process normal mixture model

9.13.1 Variable analysis on factors

The correlation for each variable against the four factors of the TrDPNMLVM can be seen in table 9.13.1 this is similar to the factor loading table from table 2 for the normally distributed factor analysis, suggesting that the factors are similar but not exactly the same for each analysis. The annotation are the same for each of the factors Factor 1 corresponded to atopy, factor 2 to eosinophilic inflammation, factor 3 to modified JACS symptom score and atopic status of dust mite and grass pollen and factor 4 to BMI and neutrophilic inflammation. With the factors established we now turn our attention to the clusters in each factor for the new truncated analysis approach.

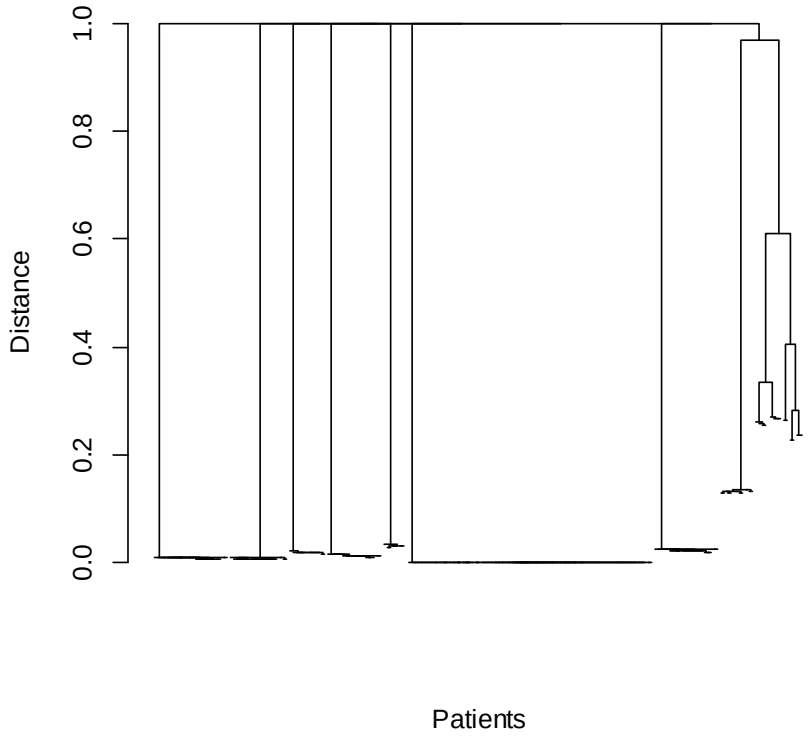
Table 9.13.1: The correlation of the variables with the factors in the truncated Dirichlet Process Normal Mixture Latent variable model, the highest factor loading for each variable is highlighted

Variables	1	2	3	4
BMI, kg/m^2	-0.08	0.01	0.03	0.78
Age of Onset of Symptoms, years	-0.38	0.12	0.36	-0.06
Log eosinophil cell count, log %	-0.07	0.99	-0.04	-0.08
Modified JACS symptom score	-0.06	-0.08	0.48	0.09
Log neutrophil cell count, log %	-0.01	-0.09	0.08	0.66
Cat skin prick test, mm	1	-0.06	-0.03	-0.03
Dog skin prick test, mm	0.83	0.08	-0.34	0.04
House Dust Mite skin prick test, mm	0.50	-0.07	-0.67	0.02
Grass Pollen Skin Prick Test, mm	0.71	0.04	-0.50	-0.01

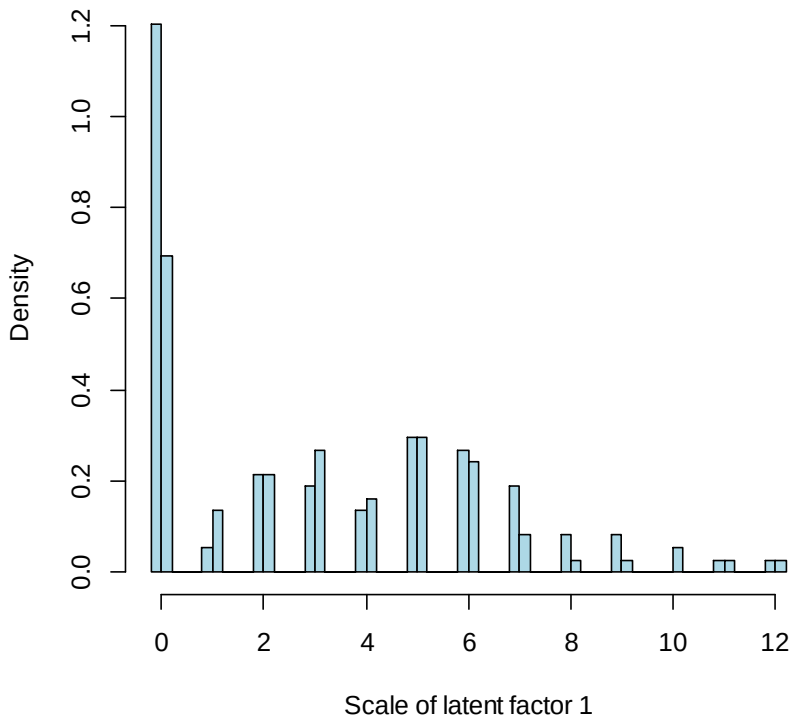
9.13.2 Factor 1: Atopy

The first factor had cluster dendrogram as in graph 9.13.2.1 suggesting the factor is separated into many clusters. Inspection of the density distribution for factor 1, See graph 9.13.2.2, suggests that the distribution has many sub-groups which are very separate but this is because the correlated atopic variables were discrete. Assuming the atopic variables were discrete was fine and produced a latent variable that was continuous in the previous analysis making the clustering valid. But in the truncated model the latent variable is discrete similar to the manifest variables and the many clusters detected represent the discrete nature of the latent variables and not true clusters.

Graph 9.13.2.1: Cluster dendrogram for factor 1 using the probability of being in a cluster with another patient to separate patients.



Graph 9.13.2.2 Histogram of the density of latent factor 1 (atopy)

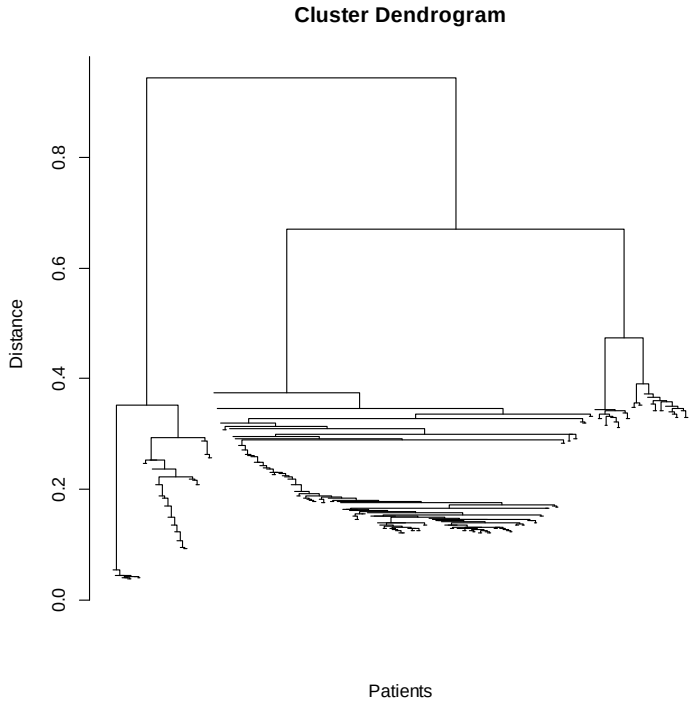


9.13.3 Factor 2: Eosinophilic inflammation

Two clusters were detected in the second factor as can be seen in the cluster dendrogram, see graph 9.13.3.1. On inspection of the density distribution for factor 2, see graph 9.13.3.2 however suggests that two groups are not easily visible, and that the two groups could make up a non-normal continuous distribution that can be approximated using two normal distributions. The dip statistic which although being significant for this factor density distribution was only just over the value at p=0.05 having dip statistic, 0.0182 (p=0.05), so if using p=0.05 as a cut off factor 2 is multi modal but not by a large degree. The only variable that was significantly associated with the eosinophilic factor other than the factor itself was BMI and age, see table 9.13.3.1. The clustering suggested two clusters one that consisted of younger patients that have lower eosinophilic inflammation in sputum and a lower BMI and the other

that consisted of a higher eosinophilia group that were older and slightly more obese which had a slightly higher nitric oxide levels in breathe.

Graph 9.13.3.1: Cluster dendrogram for factor 2 (eosinophilic inflammation) using the probability of being in a cluster with another patient to separate patients.



Graph 9.13.3.2: Histogram of the density of latent factor 2 (eosinophilic inflammation)

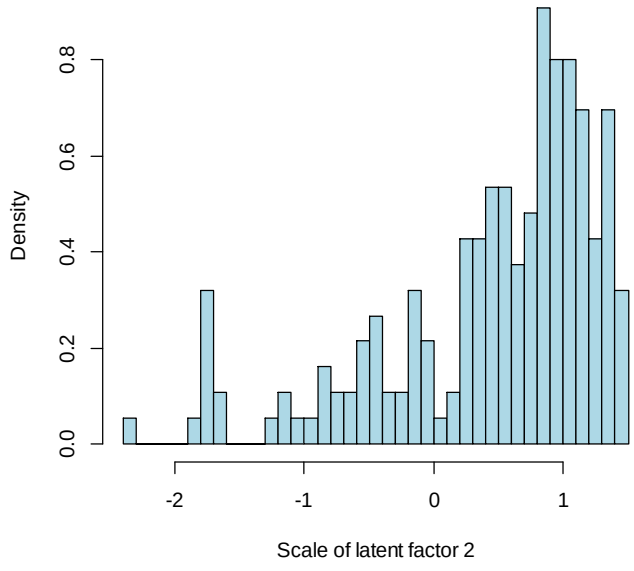


Table 9.13.3.1: The significant variables for the clusters found in factor 2 (eosinophilic inflammation)

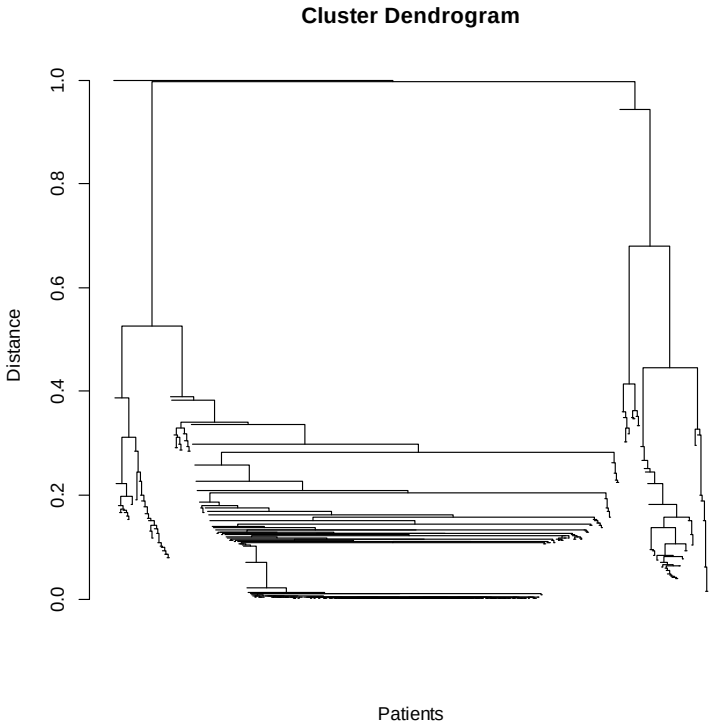
	Cluster 1	Cluster 2	p-value
N	154	33	
Age*	45 (16)	38 (15)	0.044
BMI*	28.81 (6.35)	27.14 (7.15)	0.026
Log eosinophil % in sputum*	0.82 (0.62)	-1.21 (0.66)	0.047
Log(no)*	1.70 (0.32)	1.32 (0.37)	0.001

*p-value derived from one way ANOVA

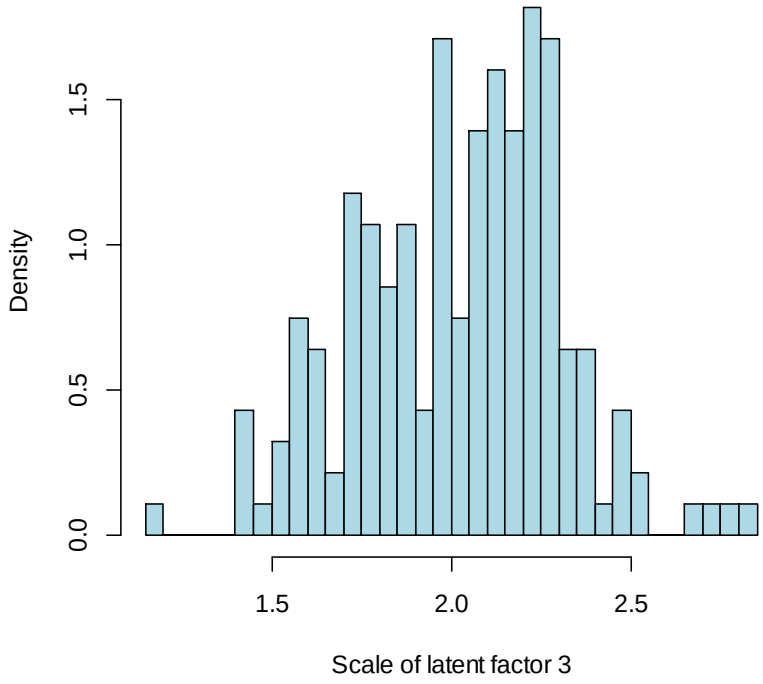
9.13.4 Factor 3: Symptoms

The third factor had cluster dendrogram as in graph 9.13.4.1 however the most probable number of subgroups that was returned by the Truncated Dirichlet process normal mixture model was only one group. Inspection of the density distribution of factor 3, see graph, 9.13.4.2 suggests that this is plausible as the distribution looks normally distributed. The evidence for no sub-groups in this factor is further backed up by the dip test being non-significant for this factors distribution, dip statistic = 0.0201 (p=0.5).

Graph 9.13.4.1: Cluster dendrogram for factor 3 (Symptoms) using the probability of being in a cluster with another patient to separate patients.



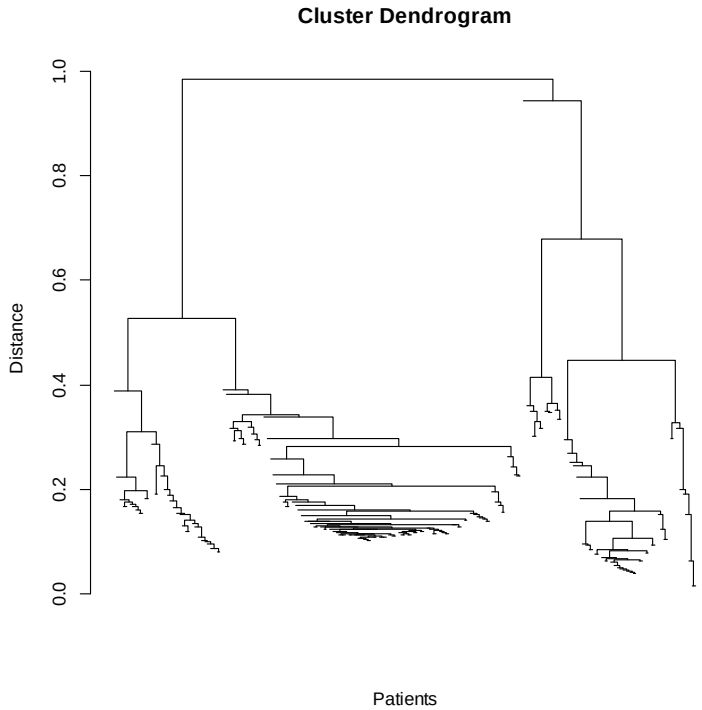
Graph 9.13.4.2: Histogram of the Density of latent factor 3 (symptoms)



9.13.5 Factor 4: BMI/neutrophils

The fourth factor had 3 prominent clusters in its cluster dendrogram, see graph 9.13.5.1, this implied the data was separated into three groups, but inspection of the density distribution of factor 4, see graph 9.13.5.2, suggests that these three groups could be making up a non-normal continuous distribution that can be approximated using the clusters. However on applying the dip statistic being non-significant for the factor density distribution, dip statistic = 0.0151, (p=0.01) confirming that the data is multi modal. One cluster only consists of one patient that is an outlier not detected in the previous analysis. The two other clusters consist of a higher neutrophilic group and a lower neutrophilic group however BMI was not significantly different over the clusters which suggests that the factor could be representing neutrophilic variation more than BMI variation, see table 9.13.5.1.

Graph 9.13.5.1: Cluster dendrogram for factor 4 (BMI/ neutrophils) using the probability of being in a cluster with another patient to separate patients.



Graph 9.13.5.1: Histogram of the density of latent factor 3 (BMI)

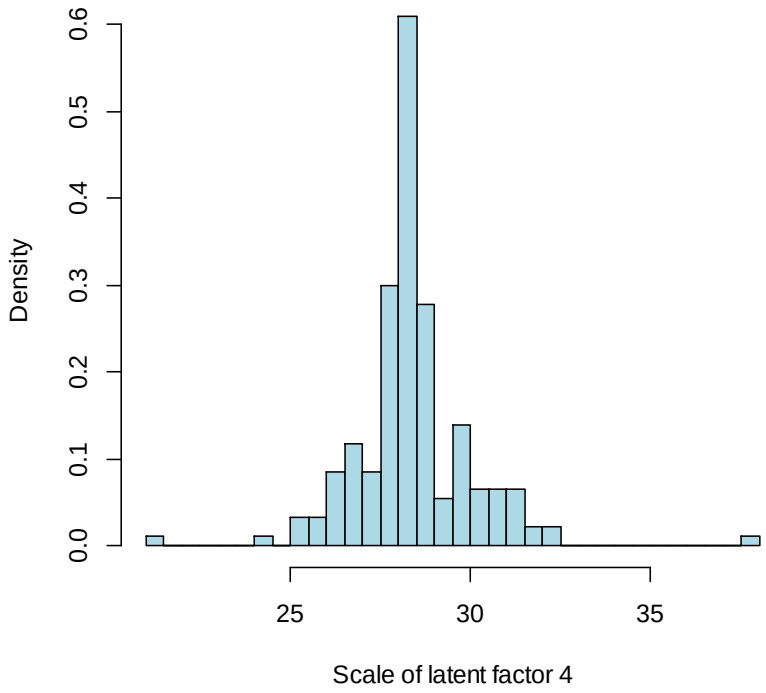


Table 9.13.5.1: The significant variables for the clusters found in factor 4 (BMI)

	Cluster 1	Cluster 2	Cluster 3	p-value
N	131	55	1	
Log neutrophil % in sputum*	1.83 (0.10)	1.32 (2.55)	-0.30 ()	0.001

*p-value derived from ANOVA test

9.14 Truncated Dirichlet process normal mixture with a binary outcome.

The binary outcome of gender was added to the model using a logit distribution and having the four latent variables driving the odds of being male/female. This model when implemented in WinBUGS could not be computed as traps were reported after a few iterations. Traps are usually entered in WinBUGS as the result of an ill specified

model. The fault could be in two parts the specifying of the priors or the specifying of the model. A range of priors and starting values were tried but were all unsuccessful. Suggesting that the specified model was perhaps to blame. The model relies on the log odds of the binary variable to be correlated to one of the four latent variables in the model of some degree, but if this was not the case the model could enter traps. All the clusters over the factors were not significant for gender suggesting that the factors in the truncated Dirichlet model did not correlate well with gender. Thus I think this is the reason the truncated Dirichlet process normal mixture model failed when a binary variable was added as the factors created previously do not correlate with gender. We could have used a five factor model but all this would have done is predict gender and we have found that there is not enough information to generate a consistent latent factor for one binary variable in our previous simulations.

9.15 Discussion

9.15.1 Dirichlet Process Normal Mixture Model

Clusters were found on all four factors with significant similar clustering seen over all three priors used on the parameter α. Although the clusters seen for each factor differed when comparing the means of the groups for asthma variables that were correlated to the individual factors, but not all factors were significantly multimodal indicating overlapping clusters creating non-normal distributions. The cluster partitions for all factors, apart from the factor relating to atopy, were significant for gender. Indicating that gender is very important for separating the clusters even though it was not used in the DPNMLVM analysis due to it being a binary variable and it not being a result of the asthma condition.

Factor 1 the atopic factor grouping suggested an early onset group that were younger and had a higher skin prick test measurement for all the triggers contrasting with a late onset older group that were less atopic to the same triggers this splitting was also found in the (Haldar,2007) and (Wenzel,2009) papers and reflects a genuine split as the dip statistic was not significant indicating a genuine multimodal distribution.

The other significant variable splitting, multimodal distribution was on the BMI factor which suggested two groups an obese group contrasting with a non-obese group. On inspection of the distribution however the two groups form a continuous non-normal distribution and not a multi-modal, well separated grouping which suggests that the BMI cluster split is artificial and does not consist of two subgroups with genuine differences, but rather one continuous distribution this is further confirmed by the dip test being significant indicating a uni-modal distribution is present for the BMI factor. The Haldar paper found a BMI splitting but did not investigate whether the splitting represented a non-normal distribution as the clustering was across many variables and was not represented as a single distribution.

The two factors representing JACS symptom score and eosinophils did not produce significant differences other than gender differences. As gender was used in the initial clustering by Haldar it could have over emphasised the partitions of JACS and inflammation as the gender variable is binary and k-means/ hierarchical clustering should only be carried out on continuous variables unless using specialist algorithms built to define the different characteristics of binary variables.

9.15.2 Truncated Dirichlet Process Normal Mixture Model comparison

The Truncated Dirichlet Process Normal Mixture Model showed a similar factor analysis to the original Dirichlet Process Normal Mixture Model and classical factor analysis but this time the emphasis on the factor 4 was more on the log neutrophil % sputum rather than BMI. The first factor in the analysis atopy had a large amount of sub-groups found across it but on inspection this was driven by the discrete nature of the variable that correlated to it causing what looked like sub-groups. Factor 2 , the eosinophilic cluster showed two clusters, one showing a group with high eosinophils and one which had low eosinophils but these were not associated with gender like in the previous full Dirichlet methodology but were associated with BMI and age with the younger, smaller BMI group being a less eosinophilic group. These grouping also had a significant dip test so can be seen as multimodal clusters. Factor 3 displayed what looks like a normal distribution and only one sub-group was detected over the whole distribution. Factor 4 showed three clusters but one of these was an outlier being in its own cluster. The two other clusters were significantly different for log neutrophil % in sputum but not BMI suggesting that the factor described neutrophils more than BMI, the distribution also tested significant for the diptest for multi-modality which it did not do for the full Dirichlet process model suggesting that the fourth factor may have picked up on a slightly different variation pattern which had more emphasis on log percentage sputum neutrophil variation than BMI variation. This suggests that although the factor analysis were both similar they have converged on slightly different solutions to the same problem which can be the case in these types of complex models that allow for the variation of multi-variables. The binary variable gender was added into the factor analysis using a Bernoulli distribution but the model

could not be run as the model entered traps. Several priors and starting values were used in order to get around the trap but none were sufficient the reason for this was that the truncated Dirichlet model factors were not strongly correlated to the binary variable so a solution could not be obtained.

9.16 Closing Statement

We have seen that the Dirichlet process returns clusters that have been detected previously with different methodology more so we have investigated the nature of these clusters to determine if they make up a non-normal distribution or come from genuine different populations. In the analysis the atopic early onset/ non atopic late onset clusters were found in the Full Dirichlet process normal mixture model to represent genuine differences but the Obese/ Non Obese groups were actually just a false separation of the data. In the Truncated Dirichlet process however clusters were found on the second factor relating to eosinophilic variation in the full Dirichlet model this was nearly significant on the dip test but in the truncated model this was significant, showing two clusters one with high eosinophils and one with low eosinophils. Clusters were also found on the fourth factor these related to neutrophilic inflammation and did not relate to BMI suggesting that the fourth factor may have picked up on different variation in the datasets with the emphasis on BMI in the full Dirichlet model and the emphasis on neutrophilic variation in the truncated Dirichlet process.

Chapter 10. Analysis of a Clinical Trial Dataset for a New Cancer Drug

10.1 Chapter overview

Here the Dirichlet Process Normal Mixture Latent Variable Model DPNMLVM has been adapted for non-conjugate data by using a truncated Dirichlet process normal mixture in WinBUGS. This new model has been used to determine clusters in a clinical trial using time to event and binary variables over a single latent variable to determine clusters that were shown to have different proportions of biomarkers and survival outcomes.

10.2 Introduction

Here we apply the Dirichlet process normal mixture methodology to a clinical trial dataset in order to determine if sub-groups can be determined within clinical trial arms and to see if these sub-groups represent specific biomarkers. The clinical trial has three outcomes two that are time to event data and one that is binary. The three correlated outcomes in the model were used to determine if there are possible sub-groups within each of the treatment arms of the clinical trial. We can adapt the DPNMLVM to determine subgroups that were previously used for the normally distributed severe asthma outcomes using the full DPNMLVM. The hope is that the methodology described could hopefully be used to determine future sub-groups in clinical trials for personalised medicine applications.

Unfortunately the full Dirichlet process with the Escobar and West, 1995 algorithm cannot be used here as the three outcomes cannot be coerced into normally distributed variables. To get around this issue I used the truncated or approximated

Dirichlet Process Normal Mixture Latent variable Model trDPNMLVM introduced previously, see chapter 5. If we view the Dirichlet process mixture as a stick breaking prior we can view the trDPNMLVM as an approximation to the full DPNMLVM where the sticks are limited to a certain number in the truncated approach and where there are infinitely many in the full DPNMLVM approach. Even though an approximation the trDPNMLVM has many advantages over the full DPNMLVM these include:

- It can be implemented in WinBUGS, free Bayesian software
- It can cope with non conjugatcy issues, as WinBUGS does the calculation of the posterior.
- It is quicker as WinBUGS is based on compiled language which is faster.
- And it can give a better understanding of the mixtures as it is based on a finite amount of mixtures rather than the infinite amount of mixtures used in the full DPNMLVM as will be demonstrated later.

10.3 The Dirichlet process as a stick breaking prior recap

We have seen previously that applying a DP to an outcome, the variable is given a non-parametric distribution consisting of a base prior distribution G_o and a concentration parameter α that describes how much faith we have in this prior. We write this as,

$$G \sim D(G_o, \alpha) \qquad \textit{Equation 210}$$

A Dirichlet process is a discrete distribution over an infinite number of real points and the probabilities associated with these points can be constructed by a stick breaking process. Imagine a stick of unit length, we break a piece off the stick s_1 and assign it to be the probability p_1 of the point y_1, the remainder of the stick having magnitude *(1-*

s_1). Then break another piece of the rest of the stick s_2 to create the probability p_2 belonging to x_2, $p_2=(1-s_1)s_2$ and so on, the last remaining part could be infinitely small, an infinite amount of points are used. The $1-s_i$ fraction after each break can be shown to have expectation $\alpha/(\alpha+1)$ with the expectation after N points being shown in the equation below (Ohlssen, Sharples et al. 2007).

$$E\left(1-\sum_{i=1}^{N-1} p_i\right) = \left(\frac{\alpha}{\alpha+1}\right)^{N-1} \qquad \textit{Equation 210}$$

The main limitation of a Dirichlet Process is that it assumes the variable can be described using a discrete probability distribution. Using the stick breaking process described above and the parameter definitions we can describe the probability density function of G as

$$G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i} \quad where\ \theta_i \sim G_o, \qquad \textit{Equation 211}$$

$$\delta_\theta\ indicates\ a\ point\ mass\ at\ \theta$$

$$Gand\ p_k = s_k \prod_{k<j} (1-s_j) \qquad \textit{Equation 212}$$

$$where\ s_j \sim Beta(1,\alpha) \qquad \textit{Equation 213}$$

If the data cannot be described using discrete probability distributions, the Dirichlet Process can be made continuous by allowing the distribution to become an infinite mixture of normal distributions (Escobar and West 1995) called a Dirichlet Process Normal Mixture. The DP is now the distribution of the mean and variance of the

components of the mixture and the individual observations are assigned to the most appropriate of these components in this special case the model becomes

$$G = \sum_{k=1}^{\infty} p_k \, N(\mu_k, \sigma_k) \quad where \ (\mu_k, \sigma_k) \sim G_o$$

Equation 214

10.4 Truncated Dirichlet Process Normal Mixtures

Dirichlet Processes and Dirichlet Process Mixtures are hard to implement due to their non-parametric nature and their non-finite number of components, but this was achieved previously by applying the Escobar and west, 1995 MCMC technique to get around these problems, but even with this methods to implement a full Dirichlet Process or Dirichlet Process Mixture is non-trivial and Dirichlet Process and Dirichlet Process Mixtures cannot be implemented in standard Bayesian software such as WinBUGS. To get around this difficulty, a truncated Dirichlet Process/Dirichlet Process Mixture model was suggested by (Ishwaran and James 2002) and has since been implemented in WinBUGS (Ohlssen, Sharples et al. 2007) although not for use with latent variables. The truncated Dirichlet Process Mixture is similar to a full Dirichlet Process mixture except that the maximum number of distributions (N) in G_o is fixed in advance. This model can be thought of as limiting the number of breaks in the stick breaking process to N-1, the last part of the stick being equal to 1 minus the other parts of the stick.

$$\sum_{i=1}^{\infty} p_i \delta_{\theta_i} \approx \sum_{i=1}^{N} p_i \, \delta_{\theta_i}$$

Equation 215

$$\sum_{k=1}^{\infty} p_k \, N(\mu_k, \sigma_k) \approx \sum_{k=1}^{N} p_k \, N(\mu_k, \sigma_k)$$ *Equation 216*

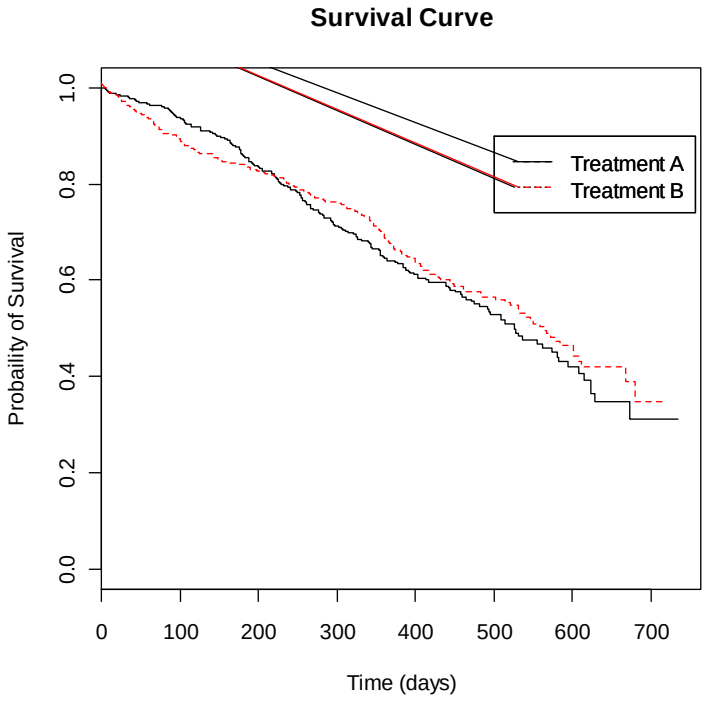
Although setting boundaries on the number of distributions is a limitation, the maximum number of groups allowed can be set to be much higher than is thought likely to occur. This is not as limiting as the k-means or finite mixture models in which the actual number of clusters has to be specified. (Ishwaran and James 2002) calculated that the L1 error bound for this approximation is

$$\approx 4\mathrm{n} \, . \, \mathrm{e}^{(-(\mathrm{N}-1)/\alpha)}$$ *Equation 217*

10.5 Analysis of a trial of a new cancer drug

A randomised clinical trial was carried out on a new anti-cancer drug to obtain licensing. Progression free times were recorded as a primary outcome along with secondary outcomes in the form of a classification of the patients as responders or non-responders and survival times. The trial recruited 1180 patients with 577 being given treatment A (chemotherapy) and 603 being given treatment B (new cancer drug). As well as outcome data the patients also have baseline data including gender, age, stage of disease, performance status and smoking status. In addition to the baseline data results for three biomarkers were obtained on a subset of patients from each trial. Gender, age and performance status were used as covariates as these were found to be significantly associated with time to progression and survival time in an initial exploration of the data. Survival curves for both treatments can be seen in Graph 10.5.1.

Graph 10.5.1 Survival curves for treatment A (Chemotherapy) and B (New Cancer Drug). Demonstrating that the two curves cross and thus making it difficult to compare the two possibly due to subgroups.



These, time to event, outcomes were modelled using a Weibull distributions with the latent variable Z driving the scale parameters. The binary outcome was modelled using a Bernoulli distribution using a logit link to connect the latent variable.

$$Y_{i1} \sim Weibull(K_1, \theta_{i1})$$ *Equation 218*

$$\theta_{i1} = log(\beta_{10} + \beta_{11} Z_i + \sum_{k=1}^{K} \gamma_{1k} X_{ik})$$ *Equation 219*

$$Y_{i2} \sim Weibull(K_2, \theta_{i2})$$ *Equation 220*

$$\theta_{i2} = log(\beta_{20} + \beta_{21} Z_i + \sum_{k=1}^{K} \gamma_{2k} X_{ik})$$ *Equation 221*

$$Y_{i3} \sim bernoulli(\theta_{i3})$$ *Equation 222*

$$\theta_{i3} = logit(\beta_{30} + \beta_{31} Z_i + \sum_{k=1}^{K} \gamma_{3k} X_{ik})$$ *Equation 220*

$$Z \sim D_{Approx}(\alpha, G_0)$$ *Equation 223*

$$G_0 \sim (\mu_k , \sigma_k)$$ *Equation 224*

Where Y_1 is time to progression, Y_2 is survival time. K_j and θ_{ij}, j=1,2, are the shape and scale parameters of Weibull distributions. Y_3 denotes whether the patient is a responder to their treatment and θ_{i3} is the probability of being a responder, Z is the latent variable driving all three outcomes, X_{ik} are covariates including the treatment indicator and β_{ij}, γ_{jk} are regression coefficients specific to variable j. Z is a mixture of normal distributions with a DP on the means/variances so the prior G_0 is over these means/precisions, a product of normal and gamma with N being the maximum number of mixtures being 15 in this case.

The progression and survival times both contained censored data. For the progression times there were 80 censored values for treatment A and 150 censored values for treatment B. For the survival outcome there were 350 censored values for treatment A and 380 censored values for treatment B. If a measurement was censored it was treated as a missing value in the WinBUGS code with a truncated range. The lower value of the range was the censored value and the higher value was kept at 2400 days as a cap as to not let extreme high values be taken as these produced traps in WinBUGS. The parameters β_{21} and β_{22} were highly correlated with each other, a familiar problem with latent variable models so these parameters were blocked together when updating to allow for the correlation seen. This was carried out by giving the two parameters a bivariate normal distribution with mean = (0, 0) and precision = $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$. These are the parameters mu_3 and pre in the WinBUGS code.

Many equivalent solutions exist for the model in order to detect just one of these β_{11} and β_{12} were both kept as the constants 0 and 1 respectively in order to obtain only one of the many transformed solutions. Mu_3, pre, β_{21} and β_{22} were all pre-specified in

the data given to the WinBUGS code along with initial starting values and do not appear in the WinBUGS code found in Appendix 8.

10.6 Priors for Model

Priors for the gamma parameters, the coefficients of the covariates were given a normally distributed prior with a mean of 0 and a precision of 0.1. The same prior was also given for the beta parameters that are associated with the intercept and the coefficient of the latent variable. The shape parameters of the two Weibull distributions were given an exponential prior with rate parameter equal to 0.5 for both, allowing a realistic range of values for the shapes and preserving congruency. The intercept and latent variable coefficient for the survival outcome were given a bivariate normal prior to allow for the correlation between the two variables this prior had a [0,0] mean and a precision of $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

The hyper-prior,mu_2 over the mean of the normal distribution parameter and mu_1 of the Dirichlet process mixture were given non-informative priors as these can influence clustering (Ishwaran and James 2002). The alpha prior was kept as a uniform distribution from 0.3 to 7 for identification purposes as suggested in (Ohlssen, Sharples et al. 2007). The maximum number of sub- groups was kept at 15 for each treatment group to obtain a fair approximation of the full Dirichlet Process Mixture. Previous models with a higher number of subgroups were used and were found to contain many empty groups.

10.7 Results

The model was fitted to all 1180 patients by including a treatment effect as one of the covariates. In a separate analysis the subjects were stratified by treatment group and the model was fitted separately to each treatment group. All three models allow for an interaction between the pattern of individual effects and the treatment. Each of the three analyses was also adjusted for age, gender and unfit at baseline. Although the trial is randomised so that covariate adjustment is not necessary for estimating the average treatment effect, covariate adjustment may affect the latent variable and hence alter the clustering of subjects. Here the latent variable will pick up factors other than age, gender and unfit at baseline that could affect individual outcome.

Each MCMC analysis was based on three chains of length 70,000 following a burn-in of 30,000 iterations. Trace plots and Gelman-Rubin statistics were checked in order to establish convergence.

The coefficient of the covariate gender was found to span zero and therefore not add significant to the progression and survival outcomes but did not span zero for the binary outcome of being a responder. However for the responder outcome, unfit at baseline and treatment covariate coefficient were not significant all other covariates however were significant for the three outcomes.

The number of normal distributions in the mixtures is shown in table 10.7.1

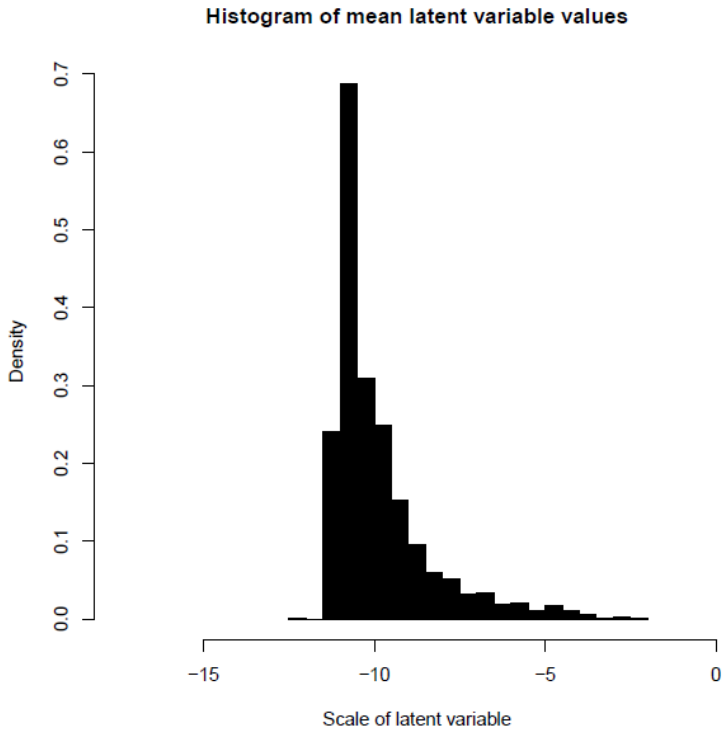
Table 10.7.1 Percentage of iterations in which the mixture involved 1 to 6+ component normal distributions

Number of Components in the mixture	All Subjects	Subjects on treatment A	Subjects on treatment B
1	0	0	0
2	0	1.46%	0
3	55.17%	50.08%	65.31%
4	30.97%	32.40%	26.53%
5	10.33%	11.66%	6.66%
6+	3.53%	4.39%	1.50%

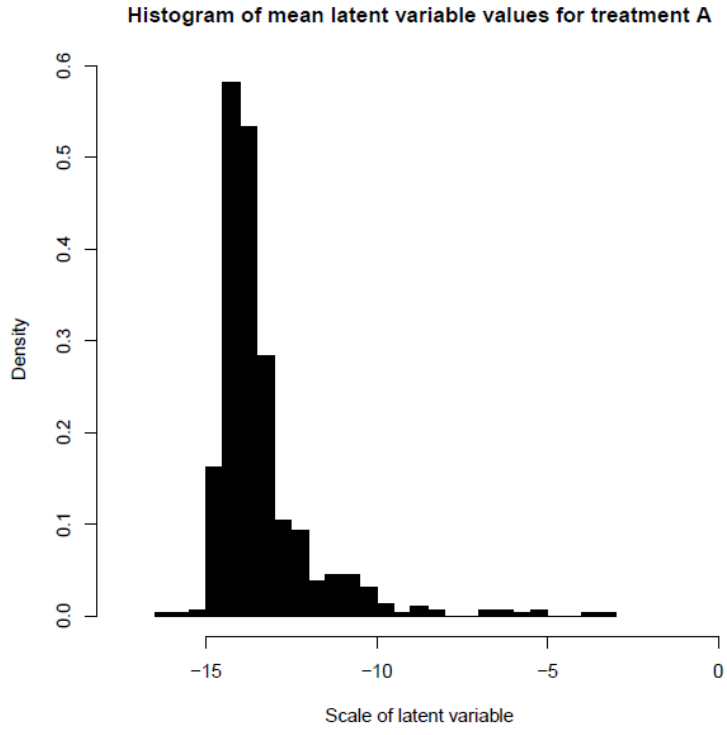
It is important not to interpret the number of components as the number of subgroups of subjects. It may be that two or more components are needed to capture a non-normal distribution of a subgroup. The most probable number of components in the mixture was three in all of the models. Graph 10.7.1 shows the distribution of the latent variable averaged over all iterations. The plots show evidence of a multi-modal distribution although this is most evident when the treatment groups are analysed separately.

Graph 10.7.1: Histogram of latent variable Z for both treatment groups combined, treatment A (chemotherapy) and treatment B (new cancer drug), each have three mixtures but these are better separated in treatment B (new cancer drug) where as in treatment A (chemotherapy) the mixtures overlap possibly into one continuous distribution made up of three mixtures rather than three separate clusters.

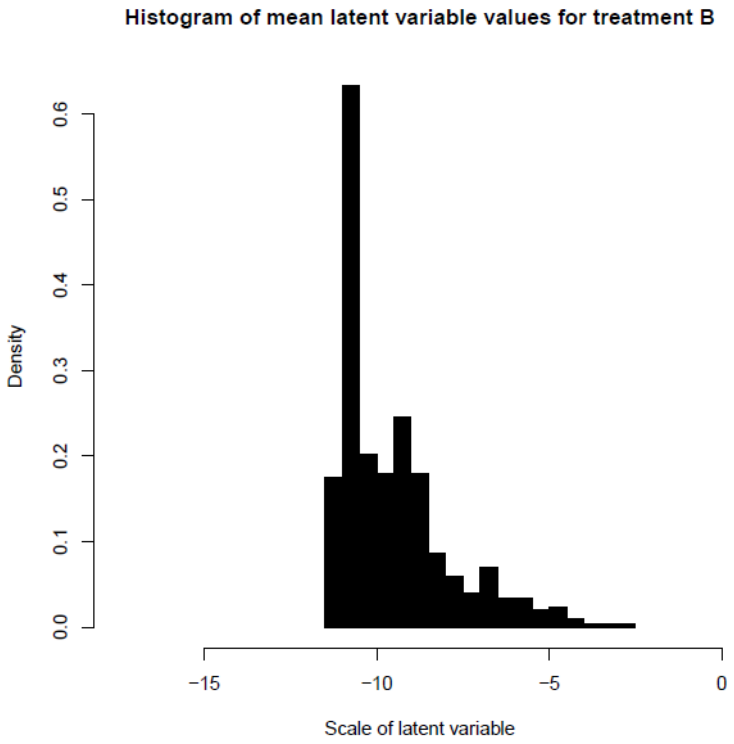
a, for all patients



b, for patients on treatment A



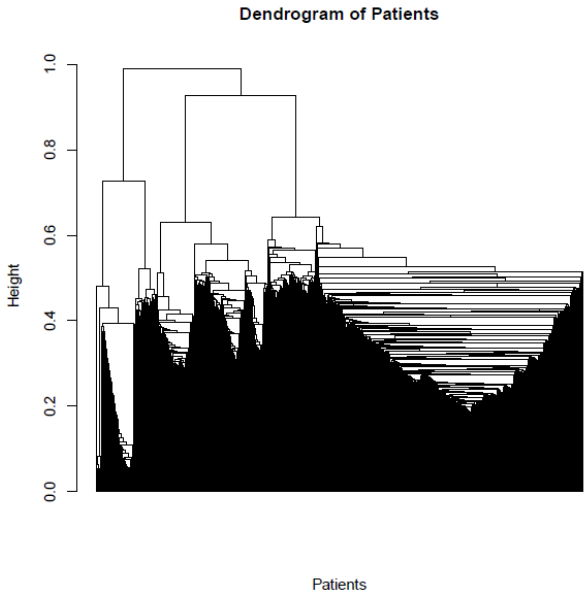
c, for patients on treatment B



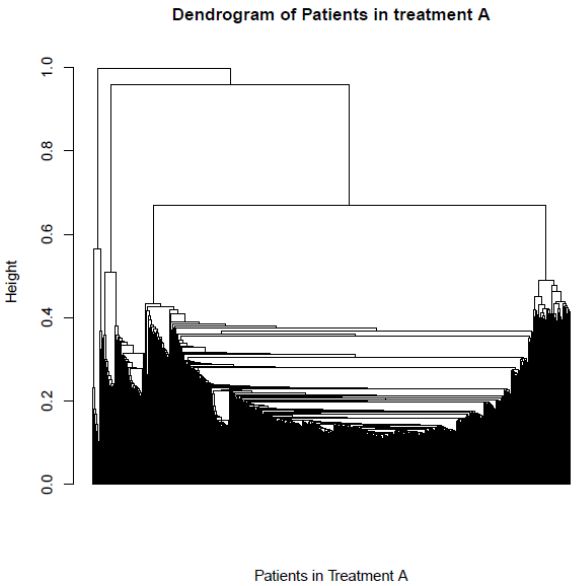
A nxn matrix was created containing the probability of belonging to the same component normal distribution for each subject pairing using all the iterations. The probability of not belonging to the same component was then used as a difference measure in a hierarchical cluster analysis giving the dendrograms shown in Graph 10.7.2, in which the solid areas are places where the tree structure is very dense. The number of clusters can then be judged by using the longest distance between branches. The best partition of the data in the hierarchical analysis was also the same as the median number of mixtures this due to the trDPNMLVM being better at determining sub-groups for large data. This was 3 in each treatment arm and when considering both modelled together with treatment as a covariate. Cutting the dendrogram at 3 clusters created a strict partition of the data.

Graph 10.7.2: Dendrogram of the clustering of subjects in both treatment groups using the probability of subject pairings not belonging to the same group as the difference measure used in the hierarchical clustering. The best number of clusters can be seen as the biggest difference between branches. This is 3 for both treatment groups.

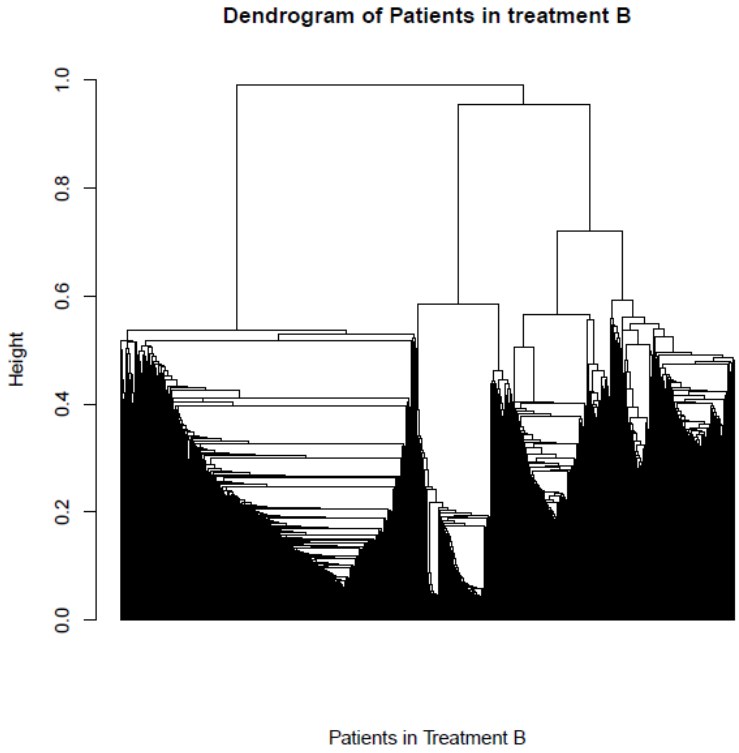
a, treatment effect model for all patients



b, for patients on treatment A



c, for patients on treatment B



Summary statistics for the mixtures can be found in table 2 for the model with both treatment groups. Summary statistics in table 3 are used for comparison of the model stratified by treatment. Statistics include results for three biomarkers that were also tested on a subset of individuals taking part in the trial. The patients tested would either have a positive or negative result for a biomarker. Percentages of patients who tested positive and negative for each mixture can also be found in tables 10.7.2 and 10.7.3.

Table 10.7.2 Summary of statistics for each subgroup in the Dirichlet Process Mixture model with added treatment effect.

	Subgroup 1	Subgroup 2	Subgroup 3	p-value
Number of subjects	763	269	148	
Time to Progression*	212(167-295)	94 (45-161)	41(24-62)	<0.001
Survival Time*	383(285-504)	264(205-340)	74(35-116)	<0.001
% Responders	55%	15%	0%	<0.001
% Males	20%	22%	21%	0.852
Age+	57 (11)	56 (11)	56 (11)	0.976
% unfit at baseline	8%	8%	21%	<0.001
% positive for Biomarker 1	73%	72%	75%	0.928
% positive for Biomarker 2	67%	54%	45%	0.003
% positive for Biomarker 3	73%	46%	20%	<0.001

* Median(IQR), p-values obtained using Kruskal Wallis test

+ mean(SD), p-values obtained using one way ANOVA

All other data was count data, p-values obtained using Chi squared test

Table 10.7.3: Summary of statistics for each subgroup in the Dirichlet Process Mixture model stratified for treatment.

		Subgroup 1	Subgroup 2	Subgroup 3	p-value
Number of subjects	A	513	50	14	NA
	B	291	217	95	NA
Time to Progression*	A	179 (130-233)	52 (42-84)	15 (7-29)	<0.001
	B	288 (210-371)	85 (43-141)	39 (23-56)	<0.001
Survival Time*	A	330 (252-448)	106 (86-157)	15 (7-29)	<0.001
	B	377 (292-521)	332 (249-428)	68 (35-123)	<0.001
% Responders	A	38.01%	2%	0%	<0.001
	B	80.8%	12.4%	0%	<0.001
% Males	A	22.42%	12.00%	7.14%	0.852
	B	20.27%	19.35%	23.16%	0.748
Age^{+}	A	56.74 (10.96)	55.96 (10.72)	56.29 (13.98)	0.884
	B	57.47 (11.27)	55.59 (11.45)	56.09 (11.34)	0.167
% unfit at baseline	A	8.38%	16.00%	42.86%	0.001
	B	28 9.62%	15 6.91%	15 6.91%	0.034
% positive for Biomarker 1	A	74.23%	76.92%	100%	0.397
	B	78.16%	60.87%	75.86%	0.193
% positive for Biomarker 2	A	61.54%	75.00%	50%	0.330
	B	77.17%	51.32%	37.84%	<0.001
% positive for Biomarker 3	A	63.21%	35.29%	50%	0.063
	B	90.72%	44.94%	10.81%	<0.001

The three groups in treatment B were found to be significant at $p<0.001$ for biomarkers 2 and 3. In treatment A however the three mixtures were not significant and seem to create a skewed normal distribution with a long right tail and it is hard to say whether the mixtures are genuine subgroups or not. The mixtures in treatment A are not significant at the 0.05 level for any biomarkers. When both treatment groups are combined and a treatment effect is added to the model biomarker 2 and 3 are still significant although it is clear from the stratified model that this difference is in treatment B the new cancer drug and not chemotherapy, treatment A.

For treatment B Group 1 survived the longest and contained mainly responders to the treatment. They also tested mainly positive for biomarkers 2 and 3. Group 2 consisted of patients who had a shorter time to progression than group 1 but with a more even split for biomarker 2 & 3; group 2 also had a lower percentage response to treatment. Group 3 had the shortest time to progression time and contained patients who were mainly negative for biomarker 2 & 3 and did not respond to the treatment. Biomarkers 2 & 3 were found to have significantly different ($p<0.001$) proportions within the 3 mixtures for treatment B.

10.8 Discussion

The main contributors to the nature of the clustering rely on the parameters of the Dirichlet process. The base distribution G_o was given non-informative hyper parameters to not influence clustering if these were changed to single values this could have an informative prior effect on the nature of the mixtures variance and means, i.e. creating a lot of clusters with small variances, this is the reason the hyper parameters were added. The alpha parameter of the Dirichlet process was kept non-informative as a uniform distribution from 0.3 to 7. The posterior of alpha has been shown to be

heavily influenced by the data with differing priors, even very conservative ones, having little effect given large amounts of data. If however small amounts of data are used the precision prior can have an effect on the clustering and the prior should be kept un-informative as either a uniform distribution or a gamma distribution of similar shape to a uniform distribution (Dorazio 2009).

The truncated Dirichlet Normal Mixture specified for WinBUGS can be modified for use with latent variables and used to find important clinical subgroups when combining multiple clinical trial outcomes. In our Cancer clinical trial data example 3 subgroups were found for the model allowing for treatment effect. However when the models were stratified and run independently for treatment group it was found that the three groups were based on different outcomes of patients on treatment B, although three mixtures were found in both treatment groups. The mixtures related to specific modes in treatment B where as in treatment A the mixtures resembled a skewed normal distribution.

The methodology presented here can be used to determine other subgroups in clinical trial data and to check biomarkers against these subgroups to determine further understanding of the disease and drug pathology and possibly an application to personalised medicine. Even if links could not be found with existing biomarkers a subgroup analysis could determine whether further investigation is warranted at phases 2 and 3 of clinical trials to obtain possible significant subgroups.

The work here extends our methodology to the new field of clinical trials and by using the truncated Dirichlet Process Normal Mixture Latent Variable Model we can adapt the methodology to non-normal outcomes and determine specific sub-groups for

these, at present however this is done by assuming that all outcome measures are correlated and can be described using only one latent variable as the methodology has not been adapted to allow multi latent variables to be used here as yet.

10.9 Closing statement

We have seen how the truncated Dirichlet Process Normal Mixture Latent Variable Model although an approximation of the full DPNMLVM is easier to program in as it is possible to implement the model in WinBUGS which automatically compute the posterior of the parameters it's self in a very fast computing language. This means that problems involving non-conjugatcy and time taken to implement can be solved easily speeding up research on the model. The approximation seems to be the best way to implement Dirichlet process/ Dirichlet process mixture models quickly.

Chapter 11. Conclusions and Further Directions

11.1 Chapter summary

Here conclusions and further directions are made both clinical and statistical in order to evaluate the work carried out and find places in both the code and the clinical decisions that could be improved and carried forward. These include adapting the model to select the best number of factors, using the mixture model to find mixtures over all factors and testing these for multi-modality and adding more variables of different types to the model. Clinical applications are to verify the clusters in bigger cohorts and apply missing data algorithms to allow for the missing data found in all cohorts and then to look at how the clusters can be implemented in severe asthma clinics.

11.2 Summary of activities

11.2.1 Statistical Work

I have programmed a Dirichlet Process Normal Mixture Model (DPNMM) and combined it with a latent variable model (LVM) to obtain a Dirichlet Process Normal Mixture Latent Variable Model (DPNMLVM). The DPNMLVM takes outcome variables that describe processes in severe asthma and separates the variables into latent variables or factors which can then be annotated to establish clinical aspects of severe asthma which are difficult to quantify, such as breathlessness. These factors also reduce the dimensionality of the data into highly correlated components. The normality assumption of the latent factors is relaxed by allowing the latent factors to be described using a Dirichlet process normal mixture prior which allows the latent variable to be made up of an infinite mixture of normal distributions. This allowed the aspects or factors of asthma to be visualised in a way

not seen before and also allowed the latent variables to be separated into distinct sub-groups or clusters.

Using the DPNMLVM it was also possible to gain inference on the sub-groups on the factors based on the probability of being in a mixture with another patient. The sub-groups making up the distribution were coupled with a statistic to determine the strength in the believability of the sub-groups this was adapted for Bayesian analysis from a frequentist statistic called the dip statistic which tests multimodality of a measurable outcomes distribution. Using the mean dip statistic it was possible to confirm if clusters found on an independent factor were genuine sub-groups or if they were part of a bigger non-normal distribution that was approximated by a mixture of normal distributions. Comparisons were made with the truncated version of the DPNMLVM to determine if similar factors and clusters could be found and if binary variables could be added in the analysis. The truncated version of the DPNMLVM was also used to determine clusters within each arm of a clinical trial based on two time to event variables and a binary variable.

11.2.2 Clinical Work

Two severe asthma datasets were analysed, the Pranab Haldar dataset and the Brompton Blood dataset. Distinct clusters were found over these dataset for the full Dirichlet model, and these were a high eosinophilic group and an average eosinophilic group in the Brompton Blood dataset and an early onset highly atopic cluster with a corresponding late onset less atopic older cluster in the Haldar dataset. These sub-groups have also been seen in the literature.

Clusters were also found that represent an obese mainly female and a non-obese grouping and also an airflow obstructed group corresponding with neutrophilic inflammation and

greater symptoms along with an opposing group that had less symptoms and a better FEV1, with less neutrophilic inflammation. Although these clusters are seen in the literature when found on independent factors it was found that they made up a larger uni-modal non-normal distribution rather than distinct clusters, although these clusters might have been under-powered. Also worth noting is that a large amount of the non-distinct clusters could be explained by gender differences, highlighting differences in asthma between male and females in this analysis. However when applying the truncated DPNMLVM the clusters found over factors were similar with two clusters being found over the eosinophilic factor, a high eosinophil group who were older and slighter more obese and a younger group that were less eosinophilic. No distinct clusters were found over JACS score in both the full and truncated Dirichlet process. The truncated Dirichlet process however found distinct clusters on factor 4 the BMI/neutrophil factor but these were significant for percentage neutrophils in sputum only. Which evidence suggests that the variation on factor 4 accounted for a larger part of the neutrophilic variation rather than the BMI variation in the truncated model. Whereas the opposite was true in the full Dirichlet model with the BMI variation being dominant over the neutrophilic variation.

In addition to the asthma datasets a clinical trial analysis was carried out using a truncated Dirichlet Process Normal Mixture Latent variable Model DPNMLVM using three clinical outcomes for a cancer clinical trial to determine the presence of sub-groups. This resulted in three distinct subgroups being found in the new drug arm of the clinical trial which had significantly different survival times and were also significant for two biomarkers.

11.3 Conclusions

11.3.1 Statistical

The DPNMLVM works well at determining the variation patterns of a large dataset, it can determine which variables are correlated with each other to make up specific factors and then determine what the underlying distribution of the factors will look like displaying these as uni-variate density distributions. Once the density distribution have been found and displayed statistics can be extracted and cluster membership can be found on each independent factor. Thus extending both factor analysis models and clustering algorithm/mixture models. Once found the clusters can be examined to determine if they make up a non-normal distribution or are in fact distinct sub-groups.

Although a good method the model is labour intensive to set up and can take a long time to converge and also suffers with difficulties in storing the results due to the large amount of space needed to store the many iterations and parameters. A better method is the truncated DPNMLVM which performs just as well despite being an approximation but can be implemented in less time due to its being able to be coded in the free Bayesian software WinBUGS or OpenBUGS. The advantages such as the automatic calculations of posterior distributions and the use of a simplified coding language and ease of using the truncated DPNMLM outweigh the disadvantages of using an approximation as specificity analysis can be carried out on datasets to determine an adequate upper limit of finite mixtures to obtain a good approximate to the infinite mixture model. Problems with both models include how many factors to determine a prior as these cannot be calculated in the model, but can be based on previous classical factor analysis. Problems in both the truncated and full DPNMLVM are that they can both produce slightly differing solutions. The solutions can be seen as comparable but may change slightly due to the variance allocation to factors within

the different models. As, if we have two manifest variables that are partially correlated and we are explaining them as a factor within a model, a choice is made in the model of how to express this common variation with how much emphasis is on each of the manifest variable variation.

11.3.2 Clinical

The model can be used to determine the patterns of variations seen when describing complex disease processes using measurable outcomes of the disease. The model determines whether the variables are correlated in processes by grouping the variables together that are similar. The model then looks to find sub-groups in the factor/process returning these clusters and determines whether the clusters are distinct sub-groups or represent non-normal severity over the factor approximated by normally distributed mixtures. The methods produced clusters and factors that have been found in the severe asthma literature confirming the model methods. The modelling concluded that we can group severe asthma into the following sub-groups which may not be mutual exclusive.

- Late onset, non-atopic, older
- Early onset, atopic, younger
- High eosinophilia
- Average eosinophilia

The following clusters have been reported in the literature but when detected using the DPNMLM were found to be a splitting up of a bigger non-normal distribution approximated by a number of normal mixtures/clusters but the diptest may have been underpowered to detect these.

- Obese, mainly female

- Obese, younger with better FEV1/FVC
- Low eosinophilic group
- High neutrophil inflammation, airway obstructed

The Truncated DPNMLM produced the following clusters

- High eosinophilia
- Average eosinophilia
- High neutrophilia
- Low neutrophilia

These clusters had a significant dip test. Many clusters were determined over the atopic factor but were found to represent the discrete nature of the manifest variables and not actual clusters like in the Full Dirichlet process which is a disadvantage of using the truncated DPNMLM as only continuous data can be treated in a continuous way and discrete variables however similar to continuous data should not be used.

11.4 Further directions

11.4.1 Statistical

The model relies on a classic factor analysis to determine the correct number of factors, this could be overcome by applying reversible jump Monte Carlo Markov Chain (MCMC) algorithms to jump between models with differing factor numbers, but this would add another layer of complexity to an already complex model, but it would be interesting to see if the reversible jump techniques could work in this case or some other method of model comparison for a Bayesian factor analysis model could be adapted for the semi-parametric factors . The factors in the model are all independent and it would be interesting to adapt the model to correlated factors and if possible to determine relationships between the

factors such as in structural equation modelling so that paths of latent variable and manifest variables could be linked together and examined for different clusters.

The model could be adapted to allow other ways of using different link functions so that any type of variables could be used not just normally distributed ones, the truncated DPNMLVM has gone part way with this by applying the techniques to both binary and time to event data, but it would be good to add count data to the list of variable types that the model could handle.

Another way of possibly extending the algorithm is to allow the clustering to be over multiple factors, but this would sacrifice the visualisation of a factor's mixtures and there is no multi-dimensional version of the dip statistic to test for multi-dimensional multimodality, which is again another area that could be looked into to adapt the dip statistic to more than one dimension.

11.4.2 Clinical

The immediate further investigation would be to carry out sub-group analysis on a larger dataset that had a larger number of patients so that the cluster factor analysis was well powered by having hopefully more than 200 patients as we have seen on smaller datasets some clusters are very close to the significant dip test and it is difficult to confirm if they are genuine clusters or not, having more patients would hopefully give a clearer indication one way or the other. Also datasets with higher number of variables and different variables to improve factor and cluster establishment and annotation would also be interesting. The inclusion of different measured variables would be a good way of investigating emerging new variables to determine whether they are measuring similar asthma pathologies as other variables or are measuring something new that could further aid diagnosis.

A key issue in large datasets such as the severe asthma datasets used here is that of missing data. In the datasets used here there was a small amount of missing data, this was dealt with by simply deleting the relevant patients form the dataset but imputation methods could be used possible incorporating the clusters to predict the missing outcome to use all of the dataset and possible obtaining a meta cluster membership from several multiple imputed datasets.

Another key future issue is what to do with the clusters once established and how can they be used in a clinical setting. They could possibly be used to predict outcomes for patients using longitudinal data or predict which treatments may be suitable for which cluster. What would be interesting is to carry out a clinical trial using a cluster related treatment and a control for each cluster to determine if such cluster related treatments could work. Another adaption could be to use the clusters as nominal traits in a genome wide association study to determine genetic biomarkers for the sub-groups given that the sub-groups used are not wholly based on environmental exposure.

References

Abraham, B., J. M. Anto, et al. (2003). "The ENFUMOSA cross-sectional European multicentre study of the clinical phenotype of chronic severe asthma." European Respiratory Journal **22**(3): 470-477.

Adcock, I. M., G. Caramori, et al. (2008). "New targets for drug development in asthma." Lancet **372**(9643): 1073-1087.

Ahmad, A. and L. Dey (2011). "A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets." Pattern Recognition Letters **32**(7): 1062-1069.

Ansari, A. and K. Jedidi (2000). "Bayesian Factor Analysis For Multilevel Binary Observations." Psychometrika **65**(4): 475-496.

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." Annals of Statistics **2**(6): 1152-1174.

Bai, T. R. and D. A. Knight (2005). "Structural changes in the airways in asthma: observations and consequences." Clinical Science **108**(6): 463-477.

Barnard, K., P. Duygulu, et al. (2003). Matching words and pictures. Workshop on Machine Learning Methods for Text and Images, Vancouver, Canada.

Bel, E. H. (2004). "Clinical phenotypes of asthma." Current Opinion in Pulmonary Medicine **10**(1): 44-50.

Berry, M., A. Morgan, et al. (2007). "Pathological features and inhaled corticosteroid response of eosinophilic and non-eosinophilic asthma." Thorax **62**(12): 1043-1049.

Bogaert, P., K. G. Tournoy, et al. (2009). "Where Asthma and Hypersensitivity Pneumonitis Meet and Differ Noneosinophilic Severe Asthma." American Journal of Pathology **174**(1): 3-13.

Braun, M., P. S. Fader, et al. (2006). "Modeling the "pseudodeductible" in insurance claims decisions." Management Science **52**(8): 1258-1272.

Brown, D. P. (2008). Efficient functional clustering of protein sequences using the Dirichlet process. Joint Meeting of the 7th European Conference on Computational Biology/5th Meeting of the Bioinformatics-Italian-Society, Cagliari, ITALY.

Burke, W. (2003). "Genomics as a probe for disease biology." New England Journal of Medicine **349**(10): 969-974.

Bush, A. and L. Fleming (1996). "Phenotypes of refractory/severe asthma." Paediatric respiratory reviews **12**(3): 177-181.

Bush, C. A. and S. N. MacEachern (1996). "A semiparametric Bayesian model for randomised block designs." Biometrika **83**(2): 275-285.

Chanez, P. and S. Wenzel (2008). "Severe asthma." Presse Medicale **37**(1): 99-105.

Chung, Y. S., D. K. Dey, et al. (2002). "Semiparametric hierarchical selection models for Bayesian meta analysis." Journal of Statistical Computation and Simulation **72**(10): 825-839.

Comrey, A. L. (1992). A first course in factor analysis L. Erlbaum Associates.

Conley, T. G., C. B. Hansen, et al. (2008). "A semi-parametric Bayesian approach to the instrumental variable problem." Journal of Econometrics **144**(1): 276-305.

Cookson, W. (2000). "Genetics of asthma and allergic disease." Human Molecular Genetics **9**(16): 2359-2364.

da Silva, A. R. F. (2007). "A Dirichlet process mixture model for brain MRI tissue classification." Medical Image Analysis **11**(2): 169-182.

Dey, D., P. Muller, et al. (1998). Practical Nonparametric and Semiparametric Bayesian Statistics. New York, Springer.

Dorazio, R. M. (2009). "On selecting a prior for the precision parameter of Dirichlet process mixture models." Journal of Statistical Planning and Inference **139**(9): 3384-3390.

Dorazio, R. M., B. Mukherjee, et al. (2008). "Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior." Biometrics **64**(2): 635-644.

Dunson, D. B. (2009). "Bayesian nonparametric hierarchical modeling." Biometrical Journal **51**(2): 273-284.

Dunteman, G. H. (1989). Principal Component Analysis. Newbury Park, California, Sage.

Entink, R. H. K., J. P. Fox, et al. (2011). "A mixture model for the joint analysis of latent developmental trajectories and survival." Statistics in Medicine **30**(18): 2310-2325.

Ernst, P., H. Ghezzo, et al. (2002). "Risk factors for bronchial hyperresponsiveness in late childhood and early adolescence." European Respiratory Journal **20**(3): 635-639.

Escobar, M. D. and M. West (1995). "BAYESIAN DENSITY-ESTIMATION AND INFERENCE USING MIXTURES." Journal of the American Statistical Association **90**(430): 577-588.

Everitt, B. (2001). Cluster analysis, Arnold, Edward.

Faffe, D. S. (2008). "Asthma: Where is it going?" Brazilian Journal of Medical and Biological Research **41**(9): 739-749.

Ferguson, T. (1973). "A bayesian analysis of some nonparametric problems." Annals of Statistics **1**: 209-230.

Folkerts, U., D. Nagel, et al. (1990). "THE USE OF CLUSTER-ANALYSIS IN CLINICAL CHEMICAL DIAGNOSIS OF LIVER-DISEASES." Journal of Clinical Chemistry and Clinical Biochemistry **28**(6): 399-406.

Gamble, J., M. Stevenson, et al. (2011). "A study of a multi-level intervention to improve non-adherence in difficult to control asthma." Respiratory Medicine **105**(9): 1308-1315.

Garcia-Aymerich, J., F. P. Gomez, et al. (2010). "Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes." Thorax **66**(5): 430-437.

Garcia-Escudero, L. A., A. Gordaliza, et al. (2010). "A review of robust clustering methods." Advances in Data Analysis and Classification **4**(2-3): 89-109.

Gelfand, A. E. and A. Kottas (2002). "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models." Journal of Computational and Graphical Statistics **11**(2): 289-305.

Gelman, A., J. Carlin, B., et al. (2004). Bayesian data Analysis, Chapman & Hall/CRC.

Ghosh, J. (2003). Bayesian nonparametrics. New York, Springer.

Guidi, J., G. Fava, A., et al. (2011). "Subtyping depression in the medically ill by cluster analysis " JOURNAL OF AFFECTIVE DISORDERS **132**(3): 383-388.

Gupta, S., S. Siddiqui, et al. (2010). "Quantitative analysis of high-resolution computed tomography scans in severe asthma subphenotypes." Thorax **65**(9): 775-781.

Halayko, A. J., T. Tran, et al. (2006). "Airway smooth muscle phenotype and function: Interactions with current asthma therapies." Current Drug Targets **7**(5): 525-540.

Haldar, P. (2008). "Cluster Analysis and Clinical Asthma Phenotypes." american journal of respiratory and critical care medicine **178**: 218-224.

Hands, S. and B. Everitt (1987). "A MONTE-CARLO STUDY OF THE RECOVERY OF CLUSTER STRUCTURE IN BINARY DATA BY HIERARCHICAL-CLUSTERING TECHNIQUES." Multivariate Behavioral Research **22**(2): 235-243.

Hartigan, J. A. and P. M. Hartigan (1985). "THE DIP TEST OF UNIMODALITY." Annals of Statistics **13**(1): 70-84.

Heidelberger, P. and P. D. Welch (1981). "A SPECTRAL METHOD FOR CONFIDENCE-INTERVAL GENERATION AND RUN LENGTH CONTROL IN SIMULATIONS." Communications of the Acm **24**(4): 233-245.

Hoff, P. D. (2005). "Subset clustering of binary sequences, with an application to genomic abnormality data." Biometrics **61**(4): 1027-1036.

Holgate, S. T. and R. Polosa (2006). "The mechanisms, diagnosis, and management of severe asthma in adults." Lancet **368**(9537): 780-793.

Hopke, P. K. and L. Kaufman (1990). "THE USE OF SAMPLING TO CLUSTER LARGE DATA SETS." Chemometrics and Intelligent Laboratory Systems **8**(2): 195-204.

Ishwaran, H. and L. F. James (2002). "Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information." Journal of Computational and Graphical Statistics **11**(3): 508-532.

Ishwaran, H. and M. Zarepour (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." Biometrika **87**(2): 371-390.

Ishwaran, H. and M. Zarepour (2002). "Dirichlet prior sieves in finite normal mixtures." Statistica Sinica **12**(3): 941-963.

Jain, S. and R. M. Neal (2004). "A split-merge Markov chain Monte Carlo procedure for the dirichlet process mixture model." Journal of Computational and Graphical Statistics **13**(1): 158-182.

Jara, A., M. J. Garcia-Zattera, et al. (2005). A Dirichlet process mixture model for the analysis of correlated binary responses. Conference on Computational Statistics and Data Analysis, Limassol, CYPRUS.

Juniper, E. F., P. M. O'Byrne, et al. (1999). "Development and validation of a questionnaire to measure asthma control." European Respiratory Journal **14**(4): 902-907.

Kaza, V., V. Bandi, et al. (2007). "Acute severe asthma: recent advances." Current Opinion in Pulmonary Medicine **13**(1): 1-7.

Kiley, J., R. Smith, et al. (2007). "Asthma phenotypes." Current Opinion in Pulmonary Medicine **13**(1): 19-23.

Kim, Y. (2003). "On the posterior consistency of mixtures of Dirichlet process priors with censored data." Scandinavian Journal of Statistics **30**(3): 535-547.

Kleinman, K. P. and J. G. Ibrahim (1998). "A semiparametric Bayesian approach to the random effects model." Biometrics **54**(3): 921-938.

Kuo, P. H., S. H. Aggen, et al. (2008). "Using a factor mixture modeling approach in alcohol dependence in a general population sample." Drug and Alcohol Dependence **98**(1-2): 105-114.

Lee, P. M. (1997). Bayesian statistics: an introduction. London, Arnold.

Lee, S. Y., B. Lu, et al. (2008). "Semiparametric Bayesian analysis of structural equation models with fixed covariates." Statistics in Medicine **27**(13): 2341-2360.

Lessard, A., H. Turcotte, et al. (2008). "Obesity and asthma - A specific phenotype?" Chest **134**(2): 317-323.

Leung, T. F., G. W. K. Wong, et al. (2005). "Clinical and atopic parameters and airway inflammatory markers in childhood asthma: a factor analysis." Thorax **60**(10): 822-826.

Lubke, H., Gita. Spies, Jeffrey, R. (2008). Choosing a 'Correct' factor mixture model. Advances in Latent Variable Mixture Models. G. Hancock, R. Chalotte, NC, Information age publishers INC**:** 343-362.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). "WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility." Statistics and Computing **10**: 325-337.

MacEachern, S. N. and P. Muller (1998). "Estimating mixture of Dirichlet process models." Journal of Computational and Graphical Statistics **7**(2): 223-238.

Maddox, L. and D. A. Schwartz (2002). "The pathophysiology of asthma." Annual Review of Medicine **53**: 477-498.

Medvedovic, M. and S. Sivaganesan (2002). "Bayesian infinite mixture model based clustering of gene expression profiles." Bioinformatics **18**(9): 1194-1206.

Milligan, G. W. (1980). "AN EXAMINATION OF THE EFFECT OF 6 TYPES OF ERROR PERTURBATION ON 15 CLUSTERING ALGORITHMS." Psychometrika **45**(3): 325-342.

Miranda, C., A. Busacker, et al. (2004). "Distinguishing severe asthma phenotypes: Role of age at onset and eosinophilic inflammation." Journal of Allergy and Clinical Immunology **113**(1): 101-108.

Mirkin, B. G. (2005). Clustering for data mining : a data recovery approach Chapman & Hall/CRC.

Moore, W. C., D. A. Meyers, et al. (2009). "Identification of Asthma Phenotypes Using Cluster Analysis in the Severe Asthma Research Program." American Journal of Respiratory and Critical Care Medicine **181**(4): 315-323.

Muller, P. and F. A. Quintana (2004). "Nonparametric Bayesian data analysis." Statistical Science **19**(1): 95-110.

Murugan, A., C. Prys-Picard, et al. (2009). "Biomarkers in asthma." Current Opinion in Pulmonary Medicine **15**(1): 12-18.

Navarro, D. J., T. L. Griffiths, et al. (2006). "Modeling individual differences using Dirichlet processes." Journal of Mathematical Psychology **50**(2): 101-122.

Ohlssen, D. I., L. D. Sharples, et al. (2007). "Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons." Statistics in Medicine **26**(9): 2088-2112.

Papaspiliopoulos, O. and G. O. Roberts (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." Biometrika **95**(1): 169-186.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Ronmark, E. (2007). "Outcome and severity of adult asthma- Report from the obstructive lung disease in northern Sweden studies (OLIN)." Respiratory Medicine **101**: 2370-2377.

Salas Hernandez, J., M. Fernandez Vega, et al. (2009). "Asthma classification." Revista alergia Mexico (Tecamachalco, Puebla, Mexico : 1993) **56 Suppl 1**: S58-63.

Skrondal, A. R.-H., Sophia. (2004). Generalized Latent Variable Modeling: Mulitilevel, Longitudinal, and Structural Equation Models. Boca Raton, Florida, CHAPMAN & HALL/CRC.

Spycher, B. D., M. Silverman, et al. (2008). "Distinguishing phenotypes of childhood wheeze and cough using latent class analysis." European Respiratory Journal **31**(5): 974-981.

Taylor, D. R., E. D. Bateman, et al. (2008). "A new perspective on concepts of asthma severity and control." European Respiratory Journal **32**(3): 545-554.

Van Hove, C. L., K. Moerloose, et al. (2008). "Cigarette smoke enhances Th-2 driven airway inflammation and delays inhalational tolerance." Respiratory Research **9**.

van Rooden, S. M., W. J. Heiser, et al. "The Identification of Parkinson's Disease Subtypes Using Cluster Analysis: A Systematic Review." Movement Disorders **25**(8): 969-978.

Ward, J., H. (1963). "Hierarchical Grouping to Optimize an Objective Function." Journal of the American Statistical Association **58**(301): 236-244.

Wardlaw, A. J., M. Silverman, et al. (2005). "Multi-dimensional phenotyping: towards a new taxonomy for airway disease." Clinical and Experimental Allergy **35**(10): 1254-1262.

Wenzel, S. (2006). "Asthma:defining of the persistent adult phenotype." The Lancet **368**(9537): 804-813.

Wenzel, S. (2006). "Physiologic and pathologic abnormalities in severe asthma." Clinics in Chest Medicine **27**(1): 29-+.

Wenzel, S. E. (2003). "A different disease, many diseases or mild asthma gone bad? Challenaes of severe asthma." European Respiratory Journal **22**(3): 397-398.

White, W., H. Johnson, et al. (2010). "Probabilistic subgroup identification using Bayesian finite mixture modelling: A case study in Parkinson's disease phenotype identification." Statistical Methods in Medical Research **0**(1).

Wu, X. D., V. Kumar, et al. (2008). "Top 10 algorithms in data mining." Knowledge and Information Systems **14**(1): 1-37.

Xing, E. P., M. I. Jordan, et al. (2007). "Bayesian haplotype inference via the Dirichlet process." Journal of Computational Biology **14**(3): 267-284.