BAYESIAN VARIABLE SELECTION IN CLUSTER ANALYSIS

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY BY RESEARCH.

By

Vasiliki Dimitrakopoulou

March 2012

Abstract

Statistical analysis of data sets of high-dimensionality has met great interest over the past years, with great applications on disciplines such as medicine, neuroscience, pattern recognition, image analysis and many others. The vast number of available variables though, contrary to the limited sample size, often mask the cluster structure of the data. It is often that some variables do not help in distinguishing the different clusters in the data; patterns over the sampled observations are, thus, usually confined to a small subset of variables. We are therefore interested in identifying the variables that best discriminate the sample, simultaneously to recovering the actual cluster structure of the objects under study. With the Markov Chain Monte Carlo methodology being widely established, we investigate the performance of the combined tasks of variable selection and clustering procedure within the Bayesian framework.

Motivated by the work of Tadesse et al. (2005), we identify the set of discriminating variables with the use of a latent vector and form the clustering procedure within the finite mixture models methodology. Using Markov chains we draw inference on, not just the set of selected variables and the cluster allocations, but also on the actual number of components, using the Reversible Jump MCMC sampler (Green, 1995) and a variation of the SAMS sampler of Dahl (2005). However, sensitivity to the hyperparameters settings of the covariance structure of the suggested model motivated our interest in an Empirical Bayes procedure to pre-specify the crucial hyperparameters. Further on addressing the problem of hyperparameters' sensitivity, we suggest several different covariance structures for the mixture components. Developing MATLAB codes for all models introduced in this thesis, we apply and compare the various models suggested on a set of simulated data, as well as on three real data sets; the iris, the crabs and the arthritis data sets.

Acknowledgments

Those pages have been the most challenging goal of my life so far. Standing at the end and looking back I can tell that it was definitely not the four easiest years in my life, but they were absolutely constructive and totally worth it.

This achievement would have not been possible without the help and guidance of my supervisors Prof. Philip J. Brown and Dr. Jim E. Griffin. I would like to sincerely thank them. Dr. Jim E. Griffin deserves a special thank you though, for all the patience, support and most importantly encouragement he gave me throughout the end of my studies. I also need to thank Prof. Petros Dellaportas for opening up the opportunity of this PhD for me and Dr. Clare Dunning for her understanding towards the end of the completion of this thesis.

Of course, all that would mean nothing and I would not be here without the support of my friends. Thank you to Dr. Jorge Nèlio Marques Ferreira for being there at the most crucial point, the very end, to hold my hand and help me give the final push for the birth of this thesis. A great thank you to still Miss, but not for long, Panagiota Maria Adamopoulou for the good laughs, stupid but hilarious emails and long phone calls that took me through the stress of writing this thesis. And finally, a special praise to my two crazy but other than that lovely people - some call them my friends too - Antonio Armando Ortiz Barrañon and Isadora Antoniano Villalobos. They were there in every single rough time I faced, to understand me, comfort me, believe in me and encourage me; but most importantly to accept me for who I am and sincerely love me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Undoubtedly, most problems of a researcher's interest have always been naturally complex and multidimensional. Forming a comprehensive image of the problem though, requires the examination of its various different aspects. However, although once manipulating such problems, and sometimes even collecting information about them, was of great difficulty, nowadays, the immense use and development of technology and computers has helped us overcome such boundaries.

Statistics, a science whose objects of study are mostly characterised by modelling complexity, has considerably contributed towards the understanding and implementation of the most elaborate problems. Certainly reinforced by the explosion of computer's usage, major development has been achieved in the area over the last decades, with multivariate statistics in particular, constituting an important tool of significant interest.

More specifically, one of the major areas of multivariate statistics is the well-known Cluster Analysis. With the primary aim of cluster analysis being the identification and construction of groups of observations that share similarities, a tremendous amount of information can be transformed and summarised into smaller segments. Benefited by a clearer condensed image of our data, patterns comprising the main source of information can therefore be identified, allowing further inference regarding the object of study. In medicine, for instance, grouping a set of patients who suffer the same disease into three different levels (e.g mild, moderate, severe), can primarily facilitate the doctor's diagnosis and indication of the required treatment to be followed, but also, looking at the problem from a different perspective, it could be used as an important tool on the determination of the causes of the disease and its level of severeness.

The beneficial features of cluster analysis have made its various techniques widely applied across many different disciplines. Besides medicine where not only patients, but also diseases or even cures themselves could be clustered, psychiatry has also made a great use of cluster analysis with clusters of symptoms, such as paranoia, schizophrenia, etc., aiding to a successful therapy. Other fields such as archeology, psychology, neuroscience, marketing, biology, machine learning, data mining, pattern recognition, image analysis, bioinformatics, climatology, crime analysis and social sciences (e.g anthropology) are only some of the many areas where clustering has been extensively used.

On the other hand, although one would think that the more information we might have the more valid the results become, analysing extended data bases is not a trivial task. In particular, in a world where everything can be monitored and measured, it is almost unavoidable that among the various variables, there is plenty of useless information or information that is replicated by many different variables. Inclusion of such unnecessary variables could not only make the interpretation of the results difficult, but also obscure the analysis itself worsening the predictions' efficiency.

Consequently, methods that can reduce the dimensions of the problem have become the focus of much research lately. Areas like text processing of internet documents, combinatorial chemistry and quality control, make great use of such methods, with gene expression array analysis and data mining holding the first places in the list.

The analysis of gene expression arrays, for example, is particularly challenging. With DNA microarrays, a multiplex technology used in molecular biology to quantify genes, being currently expensive, collection of microarray units is normally restricted to a small sample. As a result, data sets consist of an extensive number of variables - thousands of genes can be monitored in a small unit - and only a limited number of observations. The samples are commonly structured in groups that the researcher is called to recover, whereas, only subsets of genes seem to distinguish the groups. It is very important that only the really informative variables are being selected since, inclusion of non-discriminating variables can mask the group structure of the data and lead to misleading results. Models, that successfully identify the important variables and simultaneously reveal the true cluster structure, can be proved to be very useful, since a better understanding of the underlying biological complexity of the disease is being achieved and they can, consequently, serve as a powerful tool on diseases' treatment.

The microarrays' example we just presented, is a particularly representative case of simultaneously combining two major areas of multivariate statistics, the cluster analysis and the variable selection/dimension reduction, to tackle problems of biology, medicine and many other fields. To conclude though, we shall mention two more cases characterised by the special trait of many variables but just a small sample size. These are archeometry and art restoration; two very interesting areas, where only a few objects of study are available (e.g. amphorae of classical period), but with plenty of covariates available to be stored.

1.2 Preliminaries

We will now introduce essential concepts of statistics that have been used for the completion of this thesis. The following sections guide us through the most basic of those ideas, but, in the need of further definitions those should be provided at the relevant chapters.

1.2.1 Bayesian Theory

Let us begin with Bayesian statistics, an alternative to the frequentist method for statistical inference, which has gained great attention over the last decades. The originality of Bayesian statistics is hidden in the noble, whilst simple, idea of drawing conclusions about parameters of interest by using two different sources of information. The first one, as one would easily assume, is the information we gain from the observed data. Along, the second one, is the belief of the researcher himself about the parameter of interest, prior to observing the data. How these two are used in order to make inference on the parameters of interest, we will see right away.

The concept is primarily based on *Bayes' theorem*, a theorem about inverse probabilities named after the mathematician Thomas Bayes. Consider the data $x = (x_1, \dots, x_n)$ and the parameter of interest $\theta \in \Theta$, where Θ is the parameter space. The information coming prior to the observation of the data is expressed by the prior distribution $\pi(\theta)$. The data, which are independent and identically distributed (i.i.d) with distribution $p\left(x_i|\theta\right)$, give the likelihood function

$$\prod_{i=1}^{n} p\left(x_i|\theta\right).$$

According to Bayes' theorem, the likelihood updates the prior yielding the posterior distribution of θ, $\pi\left(\theta|x_1,\ldots,x_n\right)$, i.e.

$$\pi\left(\theta|x_1,\ldots,x_n\right)=\frac{\pi\left(\theta\right)\prod_{i=1}^{n} p\left(x_i|\theta\right)}{\int \pi\left(\theta\right)\prod_{i=1}^{n} p\left(x_i|\theta\right)d\theta}.$$

The parameter of interest θ now follows the posterior distribution $\pi\left(\theta|x_1,\ldots,x_n\right)$. Being interested in summaries of the posterior distribution, such as posterior mean, posterior moments, marginal densities etc, one can draw samples directly from the posterior distribution of the parameter θ, when the marginal likelihood of the model,

$$\int \pi\left(\theta\right)\prod_{i=1}^{n} p\left(x_i|\theta\right)d\theta,$$

can be calculated analytically. However, this stands for only a few cases of model-priors that result to posteriors of a conjugate form. Ergo, motivated by the need of sampling from more elaborate posterior densities, sampling methods known as Markov Chain Monte Carlo methods, have been extensively used to serve such purpose. The following section will take us through a few of the most well-known MCMC simulation methods.

1.2.2 Bayesian Inference using Markov Chain Monte Carlo methods

Markov Chain Monte Carlo methods - algorithms originally introduced by physicists in 1950's - seemed to be a promising tool for the sampling from any posterior distribution, even those of high-dimensional problems. Their engaging concept for the simulation of the parameter of interest θ drew the attention of statisticians and cast Bayesian methods on the cutting edge promoting its wide applications.

Using MCMC methods, samples for θ could now be simulated from the target distribution $\pi\left(\theta|x_1,\ldots,x_n\right)$ via Markov chains. The idea is : if we start with an arbitrary θ^0, at each step we draw θ_i's from a distribution, which as we move along the iterations converges to the stationary distribution $\pi\left(\theta|x_1,\ldots,x_n\right)$. The chain of θ_i's forms our sample for θ and inference can now be delivered.

Many different approaches have been developed within the MCMC methodology, with the Metropolis-Hastings algorithm and its variants, as well as the Gibbs sampler being the most widely used. In the following, we demonstrate the MCMC methods used for the completion of the work presented in this thesis.

The Metropolis - Hastings algorithm

As previously stated, the idea of MCMC algorithms dates back to early 50's, when Metropolis et al. (1953) introduced the Metropolis sampler as a method for sampling from the Boltzmann distribution. It was then generalised in Hastings (1970) and given the name Metropolis-Hastings algorithm, establishing it as one of the most widely used and useful MCMC samplers. Bayesian statistics was among the many fields that incorporated the use of the new sampler; a fact that contributed to the rapid evolution and extensive use of the former.

Assuming that we want to sample the parameter θ from its posterior distribution $\pi\left(\theta|x_1,\ldots,x_n\right)$, the sampler works as follows. Starting with θ^0, a value drawn say by θ's prior distribution, at any given state t of the chain with θ^t, a new θ^* is generated from a proposal distribution $q\left(\theta^*|\theta^t\right)$. The proposed value is accepted with probability

$$\alpha=\min\left\{1,\frac{\pi\left(\theta^*|x_1,\ldots,x_n\right)\cdot q\left(\theta^t|\theta^*\right)}{\pi\left(\theta^t|x_1,\ldots,x_n\right)\cdot q\left(\theta^*|\theta^t\right)}\right\},$$

and we set $\theta^{t+1}=\theta^*$, otherwise, $\theta^{t+1}=\theta^t$. The process is repeated for a certain amount of iterations, a number that is determined by the researcher according to the needs of the study, and the sampler eventually converges to the equilibrium distribution. Finally, depending on the chosen proposal, the Metropolis-Hastings algorithm has several variants, e.g the M-H independence sampler, the random walk M-H algorithm and more.

Random walk algorithm

Continuing with the random walk Metropolis-Hastings algorithm, the moves of the Markov chain for this adaptation of the M-H, are proposed such that $\theta^t=\theta^{t-1}+\epsilon$, with ϵ being a random variable the distribution of which does not depend on θ. A multivariate normal with mean 0 is extensively used as the distribution of ϵ. Ergo, with a symmetric around 0 distribution,the algorithm accepts the proposed moves with probability :

$$\alpha=\min\left\{1,\frac{\pi\left(\theta^*|x_1,\ldots,x_n\right)}{\pi\left(\theta^t|x_1,\ldots,x_n\right)}\right\}.$$

Care should be taken for the choice of the proposal distribution, as we want acceptance rates neither too low nor very high. A proposal that suggests small shifts for θ will attain a high acceptance rate as most of the moves will be accepted. However, convergence is achieved after the sampler has explored the whole parameter space. Evidently, a sampler with proposed moves close to each other will need more time to explore the parametric space and thus convergence will be achieved after many iterations. On the contrary, big moves for θ make the sampler vulnerable to low acceptance rates. That is due to the fact that many of the proposed moves are expected to lie in the tails of the target distribution and they are thus frequently rejected. More specifically, in the case of a normal proposal distribution, the distance between the moves is controlled by the variance of the increments ϵ, σ_ϵ^2. A choice of a small variance would produce values close one to each other resulting to high acceptance ratios. Meanwhile, low acceptance ratios would be drawn from a chain with large variance, for distant values would be suggested and most likely rejected. (Roberts et al., 1997) show that the optimal random walk algorithm, considering a normal proposal distribution, has acceptance rate 0.234, while on other proposals, rates between 0.1 and 0.4 perform close to optimal [Breyer and Roberts (2000) and Roberts and Rosenthal (2001)].

The Gibbs sampler

Gelfand and Smith (1990) introduced the Gibbs sampler, an idea originally proposed by Geman and Geman (1984) for applications on image processing, as a means for the sampling from complex joint posteriors. Multidimensional problems, for instance, where sampling from the joint posterior of two or more variables was once demanding, could now be availed by the new sampler.

Gibbs sampler suggests that marginalisation for the parameter of interest could be achieved, instead of integrating over the joint posterior, by sampling from the conditional distributions. Indeed, given the parameter of interest $\theta = (\theta_1, \ldots, \theta_d)$

to be sampled from the joint posterior $\pi\left(\theta|x_1,\dots,x_n\right)$, the sampler commences with a vector $\theta^0 = (\theta_1^0,\dots,\theta_d^0)$. At each state t, a new vector $\theta^t = (\theta_1^t,\dots,\theta_d^t)$ is generated by sampling each component θ_j^t from its conditional distribution and conditioned on the current values of the other parameters. To clarify, we write :

$$\begin{aligned} \theta_1^t &\sim \pi(\theta_1|\theta_2^{t-1},\theta_3^{t-1}\dots,\theta_d^{t-1},x_1,\dots,x_n) \\ \theta_2^t &\sim \pi(\theta_2|\theta_1^t,\theta_3^{t-1}\dots,\theta_d^{t-1},x_1,\dots,x_n) \\ &\vdots \\ \theta_d^t &\sim \pi(\theta_d|\theta_1^t,\theta_2^t\dots,\theta_{d-1}^t,x_1,\dots,x_n). \end{aligned}$$

Eventually, after convergence has been reached, the distribution of the chain approaches the target joint distribution, with $\theta = (\theta_1,\dots,\theta_d)$ being a draw from $\pi\left(\theta|x_1,\dots,x_n\right)$.

Likewise the M-H algorithm, several variations of Gibbs sampling can be found in the literature, e.g. the blocked Gibbs sampler and the collapsed Gibbs sampler of Liu (1994). Later in this thesis we use the Collapsed-Gibbs sampler along which a group of parameters is blocked together and sampled conditioned on all other parameters. In turn, the remaining parameters of interest are sampled having marginalised over at least one of the parameters included in the precedent group. For example, let α,β,γ be our three parameters of interest. Considering a Gibbs sampler, we would draw α,β,γ from the conditional distributions $p\left(\alpha|\beta,\gamma\right)$, $p\left(\beta|\alpha,\gamma\right)$ and $p\left(\gamma|\alpha,\beta\right)$ respectively. Under the Collapsed-Gibbs approach though, and assuming we originally decide to marginalise over β, we sample α conditioned only on γ, from the conditional distribution $p\left(\alpha|\gamma\right)$. However, for the sampling of γ, we "expand" blocking β,γ together and sampling β from $p\left(\beta|\alpha,\gamma\right)$ and γ from $p\left(\gamma|\alpha,\beta\right)$.

Gibbs sampler and its adaptations are particularly useful in the Bayesian statistics. In hierarchical models, extensively used in the applications of Bayesian statistics, the parameters of interest are conditionally independent. The full conditionals for the parameters to be sampled are then of known form, which facilitates direct sampling using the Gibbs methodology.

Reversible Jump Markov Chain Monte Carlo

On the generalisation of the Metropolis-Hastings algorithm for problems where exploring the different parametric spaces are of interest, Green (1995) introduced the Reversible Jump Markov Chain Monte Carlo sampler. With problems developed within the idea of finite mixture models being a representative example of RJMCMC being of great use, the new sampler offered the flexibility of exploring different states, allowing the number of components considered unknown.

Suggesting moves between states, RJMCMC changes the dimension of the parametric space requiring appropriate changes on the actual parameters. On that note, random variables are needed for the transformation of the parameters of interest. Assume we are in state k with parameter vector θ_k of dimension d_k. Proposing a random variable u of dimension d_u from a proposal density $q(u)$, we consider the move ℓ to state k' of dimension $d_{k'}$ and parameter vector $\theta_{k'}$. For the transformation $(\theta_k, u') = g(\theta_{k'}, u)$, an invertible and deterministic function g is needed, while dimensions need to fulfill the dimension balancing condition to secure reversibility, i.e. we need $d_k + d_{u'} = d_{k'} + d_u$, where u' is the random variable used for the performance of the reverse move. The sampler, finally, accepted move ℓ with probability :

$$\min\left\{1, \frac{\pi(\theta_{k'}|x_1, \ldots, x_n)\, r_\ell(\theta_{k'})}{\pi(\theta_k|x_1, \ldots, x_n)\, r_\ell(\theta_{k'})\, q(u)} \left|\frac{\partial(\theta_{k'}, u')}{\partial(\theta_k, u)}\right|\right\},$$

where $r_\ell(\theta_k)$ is the probability of choosing a move of type ℓ, when in state k, and $|\partial(\theta_{k'}, u') / \partial(\theta_k, u)|$ is the Jacobian of the transformation from (θ_k, u) to $(\theta_{k'}, u')$.

1.2.3 Convergence

After a certain amount of iterations, the Markov chain for the parameter of interest reaches its equilibrium distribution, also known as the algorithm has converged to the required posterior distribution, allowing the performance of inference. Averaging over the samples drawn from the posterior $\pi\left(\theta|x_1,\ldots,\_n\right)$, estimation on posterior summaries of interest can be made. However, the matter of when convergence has been achieved needs further discussion; Brooks and Roberts (1998) and Cowles and Carlin (1996) offer extensive reviews on convergence techniques for Markov Chains.

Termed as the burn-in period, a number of initial iterations are usually discarded to overcome the influence of the starting distribution. A common practice is to discard the first half of the simulated draws. However, one needs to be careful when monitoring convergence. A long enough chain is very important to ensure convergence, especially in problems with many parameters to be sampled. In such cases though, the increased amount of required storage space can be prohibitive, therefore *thinning* the chain is recommended. Under such circumstances, samples are saved only every a chosen number of iterations and inference is based on the saved draws, after convergence has been achieved.

Simulating a number of independent chains starting from several different points is also suggested. Chains centered to the same posterior distribution, with the within variation close to the between variation, suggest convergence of the algorithm. One can also calculate an estimate for the potential scale reduction (if $n \to \infty$, the factor by which the scale of the distribution of θ might be reduced). With W and B being the within and between variation of the chains respectively, we have :

$$\widehat{R} = \sqrt{\frac{\widehat{var}^{+}\left(\theta|x\right)}{W}}, \qquad (1.2.1)$$

where, $\widehat{var}^{+}(\theta|x)$ is an estimate of the marginal posterior variance of θ, such that:

$$\widehat{var}^{+}(\theta|x) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{1.2.2}$$

Finally, calculating $\widehat{R}$ for all parameters of interest, values of all $\widehat{R}$ factors close to 1 indicate convergence.

1.2.4 Distributions

Finally, at this point, we would like to introduce a couple of distributions presented in this thesis. Starting with the *Gamma* distribution, for a random variable X that follows a *Gamma* distribution with shape parameter α and scale parameter β, we write :

$$X \sim Ga(\alpha, \beta).$$

For $x \geq 0$, $\alpha \geq 0$ and $\beta \geq 0$, we define the probability density function :

$$f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

with mean α/β and variance α/β^2, where $\Gamma(\alpha)$ is the *Gamma* function. *Gamma* function is an extension of the factorial for complex and real number arguments and is defined as the integral :

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t}\, dt$$

for $R(z) > 0$. For positive integers, $\Gamma(n) = (n-1)!$.

For a random variable X, we define the Generalised Inverse Gaussian distribution over $x > 0$, with parameters $a > 0, b > 0, p \in R$ and probability density function :

$$f(x) = \frac{(a/b)^{p/2}}{2K_p\left(\sqrt{ab}\right)} x^{p-1} e^{-(ax+b/x)/2}.$$

$K_p\left(\sqrt{ab}\right)$ is the modified Bessel function of the second kind, for which

$$K_\nu(z) = \frac{1}{2}\pi\frac{I_{-\nu}(z) - I_\nu(z)}{\sin(\nu\pi)},$$

where $I_\nu(z)$ is the modified Bessel function of first kind.

1.3 Review on Variable Selection and Cluster Analysis

Earlier in this chapter we introduced the term Cluster Analysis and its extensive use in areas like medicine. We saw how such an analysis can facilitate the manipulation of vast data sets by forming groups of observations, an outcome very beneficial indeed for the interpretation and better understanding of the object under study.

For the performance of the grouping, we understand that a measurement for the indication of the observations to be clustered together is needed. A natural choice for one to make is the construction of groups of observations that are similar and close together. How is this closeness and similarity defined though? As far as closeness is concerned, a measure of distance is used to determine the observations to be clustered together. For all observations in a cluster, the chosen distance measure is meant to be small. Groups of similar observations on the other hand, are formed for observations that share a similarity measure of a large value. A variety of distance and similarity measures can be found in literature (e.g. Euclidean distance, Gower coefficient, etc), an extensive summary of which can be found in Everitt et al. (2001).

Various different clustering methods that make use of these proximity measures have been developed and used throughout the years (see Everitt et al., 2001). Hierarchical clustering, partitioning methods, with K-means being the most representative of these algorithms, and density based methods, are some of the categories lying under the wide range of Cluster Analysis.

Each of the above classical techniques, approaches the problem of clustering in a different manner. In hierarchical clustering methods, for instance, with the further categorisation in divisive and agglomerative algorithms, while groups of observations are either merged or split according to a particular distance measure, the number of clusters needs not be pre-specified. More specifically, the divisive algorithms proceed by initially considering a single cluster of observations and thereafter splitting groups that are further distant. The agglomerative algorithms on the contrary, merge groups together, starting with each observation composing a cluster. How the distance between the groups is calculated determines the agglomerative method in use (e.g. nearest/furthest neighbour, average within/between groups, centroid).

Coming to the K-means method, we first need to point-out that here, unlike hierarchical clustering, the researcher needs to predefine the number of clusters. Then, for a given number of clusters k, the algorithm starts by defining k centroids (cluster means) and calculates the distance between every observation and the k centroids. The cluster with the smallest distance from the i^{th} observation is now where this observation lies. The procedure is repeated, with the centroids being recalculated at each iteration, until the cluster allocation remains the same between two consecutive iterations.

Albeit both hierarchical and K-means clustering are simple and widely used, they come with certain disadvantages. The computationally expensive hierarchical clustering methods makes their application on large data sets prohibitive. Although K-means, on the other hand, is widely applied on large data sets working adequately and fast, it considerably depends on the initial values of the centroids, while the pre-requested information about the number of clusters is not always known to the researcher.

However, such algorithms are rather mathematical methods, that do not rely on any statistical model, but rather manipulate the data themselves overlooking the variability of the population. Statistical inference with clustering methods like K-means and hierarchical clustering, hence, cannot be performed, and therefore a method developed on the basis of a probabilistic model that adapts to our knowledge about the distribution of the data would offer more flexibility and would be more reliable for statistical inference. Also, taking into account the nature of the clustering problem and that clusters in real data sets are not particularly well-separated, we understand that uncertainty needs to be taken into account. Bayesian statistics offers such a measure of uncertainty by averaging over the posterior distribution, and thus, looking at a probabilistic model as a clustering technique within the Bayesian framework is of particular interest.

Having said that and taking advantage of the advances in computers' technology, the use of probabilistic models as a clustering method has met great success over the past years. Also known as Model-based Clustering, this method makes use of finite mixture models to model the data that are assumed as coming from G populations. The cluster assignment is then performed according to the estimated posterior probabilities (see McLachlan and Peel (2000) and Fraley and Raftery (2002) for reviews on Model-based clustering).

In the frequentists' context, clustering via finite mixture models is performed using an iterative algorithm called the EM algorithm. Considering the data

$X = (x_1, \dots, x_n)$, unobserved data Z_{ij} are introduced, where $Z_{ij} = 1$ if the i^{th} observation belongs to the j^{th} cluster and 0 otherwise, with $i = 1, \dots, n$, $j = 1, \dots, G$ and G the total number of groups. EM algorithm treats Z_{ij} as missing values and estimates the allocation probabilities w_j using the Maximum Likelihood method. The update is performed in two steps, the Expectation, where the conditional expectations of Z_{ij}, $E(Z_{ij}|X)$, are being calculated and z_{ij} are thus simply replaced by :

$$z_{ij} = \frac{w_j f\left(x_i|\theta_j\right)}{\sum\limits_{j=1}^{G} w_j f\left(x_i|\theta_j\right)},$$

and the Maximisation step, where those estimates are used to update the allocation probabilities w_j by :

$$w_j = \frac{\sum\limits_{i=1}^{n} z_{ij}}{n}.$$

The algorithm continues until $L^{k+1} \geq L^k$, where L^k is the likelihood in the k^{th} iteration.

Although, the idea of the Expectation Maximisation algorithm is simple and free from computational burdens, as it can easily be coded, there are some drawbacks as far as its convergence is concerned. Besides the very slow convergence of the EM algorithm that has been observed (see McLachlan and Peel, 2000), the relationship used as a convergence criterion, $L^{k+1} \geq L^k$, examines whether there is a change in the likelihood, rather than actually indicating convergence to a point. Moreover, a very important remark is that as the algorithm iterates, it is very likely that it gets trapped into local maxima. It is generally advised that many different starting points are used to avoid such an event, as a poor choice of starting values can lead to non-accurate estimations. In the Bayesian framework, an approach of the EM algorithm, Maximum A Posteriori EM (MAP-EM), has been developed, for which MLE is replaced by a maximum a posteriori estimator for which EM algorithm is applied. In MAP, while E-step remains the same, at step M, the conditional expectation under maximisation is augmented by the logarithm of the prior $p(.)$.

However, after the introduction of the MCMC methods in statistics, mixture model problems, and consequently Model-based clustering, are being implemented via Markov chains. With the overall population modelled using finite mixture models, under which the data are considered as coming from G different subpopulations, each of those populations follows a distribution with parameter vector θ_k and pdf $f_k(x|\theta_k)$, where k indicates the k^{th} population. We write the model ,

$$f(x) = \sum_{k=1}^{G} w_k f_k\left(x|\theta_k\right),$$

where, w_k is the weight of the k^{th} group, with $w_k \geq 0$ and $\sum_{k=1}^{G} w_k = 1$, and indicates the probability that the i^{th} observation comes from the k^{th} population. The G densities of our subpopulations can vary. However, in cluster analysis, mixtures of multivariate normal distributions are commonly chosen.

Now, instead of the classical unobserved Z_{ij}'s, within the Bayesian framework, the cluster assignment is indicated by independent latent variables y_i's. Letting $X = (x_1, \ldots, x_n)$ be independent p-dimensional observations, the y_i entries of the allocation vector $y = (y_1, \ldots, y_n)$, are assumed to be independently and identically distributed. The probability mass function of each entry is $p\left(y_i = k\right) = w_k$, where $y_i = k$ stands for the i^{th} observation coming from the k^{th} cluster. Having a prior distribution $p\left(y\right)$, we can sample the cluster indicators y_i's from their posterior distribution with the use of MCMC methods. Most commonly, a conjugate prior is used for y, allowing the sampling directly from its posterior distribution with the application of Gibbs sampler. Finally, in reality, the number of the G subpopulations is not known, while, the parametric vector θ_k is a random variable. Therefore, all G, θ_k's and weights w_k need to be estimated.

Clustering methods are, in general, a valuable tool for the manipulation and better understanding of large data sets. We saw earlier a wide range of disciplines where clustering facilitates the uncovering of possible group structures of data, with medicine being in the first line. Data sets in areas like the latter hold a special characteristic; while a vast number of variables can be collected, for various reasons, the sample size is confined.

Unraveling the group structure of high-dimensional data sets is particularly laborious. It is reasonable for one to assume that among the very many available covariates, a respectable number would be nothing more than noise, adding no extra information to the true cluster formulation of the observations. Inclusion of such noisy variables, would increase the clustering error and obscure the resulting partition. Consequently, exclusion of such unnecessary variables is of great importance.

Various methods under the general term of variable selection, have been developed to the benefit of robust models of reduced dimension. Principal component analysis is a very commonly used technique of dimension reduction, which condenses the information of the original variables in a few linear transformations, the components. Literature encompasses a variety of approaches that can reduce the dimensionality of the problem by either identifying the set of variables that best explain the partitioned space or methods that differentially weight the complete set of variables. An extensive review on variable selection methods can be found in O'Hara and Sillanpää (2009) and Dellaportas et al. (2000).

Generally speaking, the clustering procedure can proceed the variable selection; Liu et al. (2003) applied a mixture model with fixed number of clusters after a principal components analysis. However, there are cases, under which, certain variables can work as discriminating only when in conjunction with other variables. In many variable selection approaches though, such variables are considered nuisance and are excluded from the analysis (Tadesse et al., 2005). Such a result could mislead clustering accuracy. Hence, we are interested in approaches that perform the two tasks simultaneously. Methods addressing the problem of clustering and variable selection simultaneously have been developed in both the classical and Bayesian framework.

From the frequentists' point of view, one of the first attempts in the area was made by Fowlkes et al. (1988), who proposed a forward selection approach in combination with a complete linkage hierarchical clustering. Lately though, new different paths have been explored. The HINoV approach of Carmone JR et al. (1999) examines the K-means clustering in selecting variables according to certain adjusted Rand indices that pass some sort of threshold, while Brusco and Cradit (2001) perform K-means clustering in combination with a forward selection procedure. There is also a method suggested by Friedman and Meulman (2004) which combines hierarchical clustering with a weighting procedure over the complete set of variables.

While the methods above sift through the cluster assignments in the context of non-model based clustering, Raftery and Dean (2006) perform the variable selection task within model-based clustering. Adopting approximated Bayes factors, they select the "best" subset of variables by means of model comparison, with the number of clusters also estimated; a beneficial feature which is not met by methods that perform hierarchical or K-means clustering. A comparison of eight methods performing variable selection in model and non-model based cluster analysis can be found in Steinley and Brusco (2008).

In the Bayesian framework now, Heard et al. (2006) proposed a Bayesian model-based hierarchical clustering algorithm on the detection of structures within the data. Clustering genes of similar profiles can highlight possible biological mechanisms than can be further investigated. On the same direction, Medvedovic and Sivaganesan (2002) and Ramoni et al. (2002) perform clustering of gene expression profiles using Bayesian infinite mixture models.

Further on the Bayesian methodology, variable selection has been mostly developed in the regression context [George and McCulloch (1997) and Brown et al. (1998)]. A work by Tadesse et al. (2005) has, however, introduced the idea of exploring the covariates' space via a stochastic search on a latent vector-indicator of the important variables. The inclusion/exclusion search has been combined with a clustering procedure with unknown number of clusters, based on the reversible jump MCMC idea proposed by Richardson and Green (1997) and extended to the multivariate case.

Prior to the RJMCMC technique, Diebolt and Robert (1994) had explored MCMC methods, such as the Gibbs sampler, on clustering with a known number of components, while Stephens (2000) introduced a Markov birth-death process as an alternative to the reversible jump. Nobile and Fearnside (2007) proposed the allocation sampler, a new MCMC technique, that integrates out all the component parameters and samples only the number of clusters and the allocation variables.

From the nonparametric point of view, the cluster structure is explored using Dirichlet Process mixture models. On the model proposed in Tadesse et al. (2005), Kim et al. (2006) consider an infinite number of components and form the clustering task in the context of a Dirichlet Process mixture model. Hoff (2006), on the other hand, proposed a more general variable selection strategy, where different subsets of variables are identified as discriminating along the various groups according to Bayes factors. Using normal mixtures, clusters are estimated employing a mean shift approach, while MCMC methods are used to update the model parameters. Based on the mean shift model by Hoff and using a Dirichlet process shrinkage approach for the selection of variables, Lian (2010) makes his contribution to the scheme of variable selection within clustering for high-dimensional problems. On DNA microarrays, Dahl (2006) proposes a conjugate Dirichlet Process mixture model which clusters genes based on their treatment effects and variance, overcoming the need of estimating the number of clusters. Finally, Yau and Holmes (2011) propose a hierarchical Bayesian nonparametric mixture model for the performance of clustering, combined with a cluster specific variable selection.

Chapter 2

Computational Methods for Bayesian Variable Selection in Cluster Analysis

2.1 Prologue

Earlier in chapter 1, we discussed the need of introducing the variable selection methodology into exploring the cluster formulation of high-dimensional data sets and how identifying the important variables can improve the performance of the clustering task. Being interested in applying the two methodologies simultaneously, we approach both variables' identification as well as the clustering task within the Bayesian framework. Tadesse et al. (2005) incorporated model-based clustering into formulating a Bayesian model using a latent variable that determines the identification of the important variables. Using Monte Carlo methods and assuming an unknown number of components, the cluster structure of the data is being uncovered with regard to a small subset of variables, chosen as the important and discriminating ones, while the Reversible Jump MCMC technique provides the additional benefit of estimating the number of clusters.

Motivated by the work of Tadesse et al. (2005) and on their suggested model, we explore alternative computational methods. Starting, fitting the model built by Tadesse et al. (2005), we choose to alter the split/merge moves developed within the Reversible Jump technique such that moves on both empty and non-empty components can now be proposed resulting to a simpler proposal. Describing the model with its prior and posterior settings, the following sections provide the detailed steps of the MCMC methodology for the estimation of the parameters of interest, together with applications on a simulated data set. Further, investigating the sensitivity of the model to the settings of the hyperparameters, we propose a data-based method en route to the selection of the crucial hyperparameters. We designed a preprocessing procedure of two stages that precedes the analysis and takes into account the information provided from the data to anticipate values for the hyperparameters of interest; a procedure that could be considered an empirical Bayes method. Finally, we consider an alternative approach to the split/merge move, which sequentially allocates the observations with probabilities conditioned on previously allocated data (Dahl, 2005).

2.2 Model

As stated earlier, throughout this chapter we will be fitting the model as suggested by Tadesse et al. (2005) examining a few different computational approaches. The model has been built on the idea of mixture models, which delivers the clustering task, and the inclusion of a latent vector that facilitates the selection of variables. In chapter 1, we examined the clustering task formed within the mixture models framework, from both Bayesian and classical perspective, but let us now see how model-based clustering applies when in use with the variable selection task.

2.2.1 Model-Based Clustering

Let $X = (x_1, \ldots, x_n)$ be independent p-dimensional observations. Data are viewed as coming from G different populations, each represented by a distribution, while a latent vector γ of dimension p with binary entries such that :

$$\gamma_j = \begin{cases} 1, & \text{if the } j^{th} \text{ variable is discriminating} \\ 0, & \text{if the } j^{th} \text{ variable is non-discriminating,} \end{cases}$$

indicates the important variables included in the model. Indices D, ND are used to denote the set of discriminating (for which $\gamma_j = 1$) and non-discriminating variables (for which $\gamma_j = 0$) respectively. Also the total number of the discriminating variables included in the model is $p_\gamma = \sum_{j=1}^{p} \gamma_j$.

For the discriminating variables x_i^D, we form a multivariate mixture model with G components,

$$f(x_i^D | w, \theta) = \sum_{k=1}^{G} w_k f\left(x_i^D | \theta_k\right), \tag{2.2.1}$$

where, $w = (w_1, \ldots, w_G)$ are the component weights with $w_k \geq 0$ and $\sum\limits_{k=1}^{G} w_k = 1$. Note that $f\left(x_i^D | \theta_k\right)$ is the density of the i^{th} observation coming from the k^{th} cluster. For the non-discriminating ones, on the other hand, the clustering scheme is regarded as sampling from a single multivariate distribution $f\left(x_i^{ND} | \theta^*\right)$.

At this point, latent variables $y = (y_1, \ldots, y_n)$ are introduced to indicate the cluster assignment of each observation. When the i^{th} observation comes from the k^{th} cluster, we write $y_i = k$, while we assume y_i's as independently and identically distributed variables, with probability mass function $p\left(y_i = k\right) = w_k$. Considering the case of the x_i coming from the k^{th} component, being normally distributed, we have : $\quad x_i^D | y_i = k, \theta, \gamma \sim N\left(\mu_k^D, \Sigma_k^D\right), \quad x_i^{ND} | \theta^*, \gamma \sim N\left(\mu^{ND}, \Sigma^{ND}\right).$

We can now write the likelihood function as :

$$\begin{aligned} & L\left(G, \gamma, w, \mu^D, \Sigma^D, \mu^{ND}, \Sigma^{ND} | X, y\right) = \\ & (2\pi)^{-(p-p_\gamma)n/2} \left|\Sigma^{ND}\right|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i^{ND}-\mu^{ND}\right)^T \Sigma^{ND^{-1}}\left(x_i^{ND}-\mu^{ND}\right)\right\} \\ & \times \prod_{k=1}^{G}\left[(2\pi)^{-p_\gamma n_k/2}\left|\Sigma_k^D\right|^{-n_k/2} w_k^{n_k}\right] \exp\left\{-\frac{1}{2}\sum_{x_i \in C_k}\left(x_i^D-\mu_k^D\right)^T \Sigma_k^{D^{-1}}\left(x_i^D-\mu_k^D\right)\right\}, \end{aligned} \tag{2.2.2}$$

where $C_k = \{x_i | y_i = k\}$ with cardinality n_k.

2.3 Prior Settings and Full Conditionals

Continuing with the model specifications, we will now examine the prior settings and the resulting full conditionals of the model. We give details on both cases of assuming unequal covariance matrices across the groups (heterogeneous case), as well as for the case of groups sharing a single covariance matrix (homogeneous case).

2.3.1 Prior formulation

We begin by assuming that the elements, γ_j, of the latent vector γ are independent Bernoulli random variables such that, $p(\gamma_j = 1) = \phi$ and so

$$p(\gamma) = \prod_{j=1}^{p} \phi^{\gamma_j}(1-\phi)^{1-\gamma_j}. \tag{2.3.1}$$

The number of discriminating variables, p_γ, is then binomially distributed. We view ϕ as the proportion of the variables that we expect a priori to be discriminating, i.e. $E(\phi) = p_{prior}/p$, where p_{prior} is the a priori expected number of discriminating variable. Finally, by choosing a $Be(a, b)$ hyperprior on ϕ, we finally have a beta-binomial prior for p_γ with expectation $pa/(a+b)$. Setting the hyperparameters of Beta such that $a + b = 2$ we form a vague prior. Note here that for $a/(a+b) = 1/2$, i.e. $a = b = 1$, we have a uniform prior.

We set a discrete uniform prior on $[1, \ldots, G_{max}]$ as a prior on the number of components G,

$$P(G = g) = \frac{1}{G_{max}}, \tag{2.3.2}$$

while, a symmetric Dirichlet prior is chosen for the vector of component weights w, i.e. $w|G \sim Dirichlet(\alpha_w, \ldots, \alpha_w)$.

The model can be fitted using MCMC methods; however, sampling γ, as well as the number of components G as suggested by the Reversible Jump moves, changes the dimensionality of the model. As a result, parameters $\left(\mu_k^D, \mu^{ND}, \Sigma_k^D, \Sigma^{ND}\right)$ for the new components need to be sampled every time the dimension changes. Nonetheless, the increased number of parameters would affect the fitting of the model, slowing down convergence (Tadesse et al., 2005). On the matter and in favour of a more efficient algorithm, Tadesse et al. (2005) recommend the use of conjugate priors for the mean vectors and the covariance matrices to facilitate integration and therefore overcome the problem of changing dimensions. A natural assumption for one to make is that of the prior component mean being proportional to the prior component variance. We therefore have :

$$\mu_k^D|\Sigma_k^D, G \sim N\left(\mu_0^D, h_1\Sigma_k^D\right),$$
$$\mu^{ND}|\Sigma^{ND} \sim N\left(\mu_0^{ND}, h_0\Sigma^{ND}\right),$$

$$
\begin{aligned}
\Sigma_k^D | G &\sim IW\left(\delta; Q^D\right), \\
\Sigma^{ND} &\sim IW\left(\delta; Q^{ND}\right).
\end{aligned} \tag{2.3.3}
$$

Using the notation of Brown (1993), $IW\left(\delta; Q^D\right)$ represents the p_γ dimensional inverse-Wishart distribution with shape parameter $\delta = n - p_\gamma + 1$, mean $Q^D / \left(\delta - 2\right)$ and n degrees of freedom. With data matrix $X_{n \times p}$, Q^D and Q^{ND} are set as :

$$
Q^D = \frac{1}{\kappa_1} I_{p_\gamma}, \quad Q^{ND} = \frac{1}{\kappa_0} I_{p - p_\gamma},
$$

with κ_1 and κ_0 being between 1% and 10% of the upper and lower deciles of the $n - 1$ non-zero eigenvalues of the covariance matrix of the data $(Cov(X))$ respectively (Tadesse et al., 2005).

The hyperparameters of the Normal priors are chosen in favour of fairly flat priors over the variation of the data. In particular, μ_0 is simply a vector with elements the medians of the p covariates. However, there are no rules of thumb for the choice of the hyperparameters h_0 and h_1; Tadesse et al. (2005) suggest using arbitrarily large values. However, clustering is a model choice problem and so may suffer from Lindley's paradox when h_0 and h_1 are chosen arbitrarily large. Also, care needs to be taken on the choice of the constants κ_1 and κ_0, when forming matrices Q^D, Q^{ND}. We examine the sensitivity on the choice of h_1, h_0, as well as κ_1, κ_0, on our application of the model on a simulated data set at the end of the chapter.

2.3.2 Full Conditionals

The choice of conjugate priors on the component mean and variances as given in (2.3.3) allow for analytic integration of the posterior distribution over the $\mu_k^D, \Sigma_k^D, \mu^{ND}$ and Σ^{ND} parameters. The joint posterior is then of the remaining parameters (G, w, γ, y) and of the following form :

$$f\left(X, y|G, w, \gamma\right) = \int f\left(X, y|\gamma, G, w, \mu^{D}, \Sigma^{SD}, \mu^{ND}, \Sigma^{ND}\right) f\left(\mu^{D}|G, \gamma, \Sigma^{D}\right) f\left(\Sigma^{D}|G, \gamma\right)$$
$$\times f\left(\mu^{ND}|\Sigma^{ND}, \gamma\right) f\left(\Sigma^{ND}|\gamma\right) d\mu^{D}\, d\Sigma^{D}\, d\mu^{ND}\, d\Sigma^{ND}. \qquad (2.3.4)$$

Considering the cases of both equal and unequal covariance matrices, we include the calculations of the analytic integration over the parameters $\mu_k^D, \Sigma_k^D, \mu^{ND}$ and Σ^{ND} (see Appendix A).

For the case of Homogeneous covariances we obtain the following posterior distribution :

$$f\left(X, y|G, w, \gamma\right) = \pi^{-np/2} K^{D} \left|Q^{D}\right|^{(\delta+p_\gamma-1)/2} \left|Q^{D} + \sum_{k=1}^{G} S_k^{D}\right|^{-(n+\delta+p_\gamma-1)/2}$$
$$\times H^{ND} \left|Q^{ND}\right|^{(\delta+p-p_\gamma-1)/2} \left|Q^{ND} + S^{ND}\right|^{-(n+\delta+p-p_\gamma)/2}, \qquad (2.3.5)$$

with,

$$K^{D} = \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-p_\gamma/2}\right] \prod_{j=1}^{p_\gamma} \frac{\Gamma\left(\frac{1}{2}\left(n + \delta + p_\gamma - j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta + p_\gamma - j\right)\right)},$$

$$H^{ND} = \left(h_0 n + 1\right)^{-(p-p_\gamma)/2} \prod_{j=1}^{p-p_\gamma} \frac{\Gamma\left(\frac{1}{2}\left(n + \delta + p - p_\gamma - j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta + p - p_\gamma - j\right)\right)},$$

$$S_k^{D} = \frac{n_k}{h_1 n_k + 1}\left(\mu_0^D - \overline{x}_k^D\right)\left(\mu_0^D - \overline{x}_k^D\right)^T + \sum_{x_i^D \in C_k}\left(x_i^D - \overline{x}_k^D\right)\left(x_i^D - \overline{x}_k^D\right)^T,$$

$$S^{ND} = \frac{n}{h_0 n + 1}\left(\mu_0^{ND} - \overline{x}^{ND}\right)\left(\mu_0^{ND} - \overline{x}^{ND}\right)^T + \sum_{i=1}^{n}\left(x_i^{ND} - \overline{x}^{ND}\right)\left(x_i^{ND} - \overline{x}^{ND}\right)^T,$$

where, $\overline{x}_k^D$ is the sample mean of the variables included in the model for the k^{th} cluster and $\overline{x}^{ND}$ is the sample mean of the chosen as non-discriminating ones. Under the assumption of Heterogeneous covariance matrices, the joint posterior becomes :

$$f\left(X, y|G, w, \gamma\right) = \pi^{-np/2} \prod_{k=1}^{G} \left\{ K_k^D \left|Q^D\right|^{(\delta+p_\gamma-1)/2} \left|Q^D + S_k^D\right|^{-(n_k+\delta+p_\gamma-1)/2} \right\}$$

$$\times\, H^{ND} \left|Q^{ND}\right|^{(\delta+p-p_\gamma-1)/2} \left|Q^{ND} + S^{ND}\right|^{-(n+\delta+p-p_\gamma)/2}, \quad (2.3.6)$$

with, $K_k^D \;=\; w_k^{n_k} \left(h_1 n_k + 1\right)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\left(\frac{1}{2}\left(n_k + \delta + p_\gamma - j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta + p_\gamma - j\right)\right)}$, and H^{ND}, S_k^D, S^{ND} as defined earlier.

The full conditionals of the parameters y, γ and w, then become :

$$f\left(y|G, w, \gamma, X\right) \propto f\left(X, y|G, w, \gamma\right), \quad (2.3.7)$$

$$f\left(\gamma|G, w, y, X\right) \propto f\left(X, y|G, w, \gamma\right) p\left(\gamma|G\right), \quad (2.3.8)$$

where, $p\left(\gamma|G\right)$ is the beta-binomial distribution, and

$$w|G, \gamma, y, X \sim \mathit{Dirichlet}\left(\alpha_w + n_1, \ldots, \alpha_w + n_G\right). \quad (2.3.9)$$

Finally, to be more precise, we give the final form of the full conditionals in (2.3.7) and (2.3.8). For the case of homogeneity these are :

$$f\left(y|G, w, \gamma, X\right) = K^D \left|Q^D + \sum_{k=1}^{G} S_k^D\right|^{-(n+\delta+p_\gamma-1)/2},$$

$$f\left(\gamma|G, w, y, X\right) = K^D \left|Q^D\right|^{(\delta+p_\gamma-1)/2} \left|Q^D + \sum_{k=1}^{G} S_k^D\right|^{-(n+\delta+p_\gamma-1)/2}$$

$$\times\, H^{ND} \left|Q^{ND}\right|^{(\delta+p-p_\gamma-1)/2} \left|Q^{ND} + S^{ND}\right|^{-(n+\delta+p-p_\gamma)/2}$$

$$\times\, \frac{B\left(p_\gamma + a, p - p_\gamma + b\right) \binom{p}{p_\gamma}}{B\left(a, b\right)},$$

while, for heterogeneity we have :

$$\begin{aligned}
f\left(y|G,w,\gamma,X\right) &= \prod_{k=1}^{G}\left\{K_k^D\left|Q^D\right|^{(\delta+p_\gamma-1)/2}\left|Q^D+S_k^D\right|^{-(n_k+\delta+p_\gamma-1)/2}\right\},\\
f\left(\gamma|G,w,y,X\right) &= \prod_{k=1}^{G}\left\{K_k^D\left|Q^D\right|^{(\delta+p_\gamma-1)/2}\left|Q^D+S_k^D\right|^{-(n_k+\delta+p_\gamma-1)/2}\right\}\\
&\quad\times H^{ND}\left|Q^{ND}\right|^{(\delta+p-p_\gamma-1)/2}\left|Q^{ND}+S^{ND}\right|^{-(n+\delta+p-p_\gamma)/2}\\
&\quad\times\frac{B\left(p_\gamma+a,p-p_\gamma+b\right)\binom{p}{p_\gamma}}{B\left(a,b\right)}.
\end{aligned}$$

2.4 Posterior Inference

Moving now to the estimation of the parameters of interest : y,γ,w, and G, we will describe the MCMC techniques used to draw the posterior samples. Starting from the variable selection vector γ and after updating it from its full conditional in equation (2.3.8), we proceed with the parameters of the clustering task w,y, and we finish up with the split/merge and birth/death moves of the Reversible Jump procedure, that allow our search to explore different dimensional spaces by creating or deleting empty components. But let us go through each of the steps of the algorithm and examine them in detail.

2.4.1 Variable Selection

First in place comes the update of γ, the vector that will indicate which variables are chosen to be included in the model. Using a Metropolis search and three different types of moves, a new candidate γ' is proposed and then accepted or rejected according to a probability. The moves that suggest the new candidate are:

1. *Add* : Randomly choose one of the 0 elements in γ and change it to 1.

2. *Delete* : Randomly choose one of the 1 elements in γ and change it to 0.

3. *Swap* : Randomly and independently choose a 0 and a 1 in γ and switch their values.

The algorithm chooses randomly, with probability 1/3, between the three moves and accepts the new candidate with probability $\min\left[1, \frac{f(\gamma'|X,y,w,G)}{f(\gamma|X,y,w,G)}\right]$.

2.4.2 Cluster Allocation and Weights

Following the selection of variables, we come across the sampling of the two parameters that yield the cluster separation; that is, the weights w and the cluster allocation vector y. Evidently, as results from the full conditional (2.3.9), the component weights can be drawn using a Gibbs sampler. As far as the vector y is concerned, its elements get updated one at a time under a sub-Gibbs sampling strategy. More specifically, the full conditional probabilities that the i^{th} observation belongs to the k^{th} cluster are being calculated, for every cluster and for all the observations, using the cluster assignment of the remaining observations $y_{(-i)}$. The probabilities are calculated according to :

$$f\left(y_i = k|X, y_{(-i)}, \gamma, w, G\right) \propto f\left(X, y_i = k, y_{(-i)}|G, w, \gamma\right). \tag{2.4.1}$$

2.4.3 Reversible-Jump MCMC

Having sampled the new cluster assignments and component weights, we want the algorithm to further explore the dimensional space. Under the idea of the RJMCMC sampler [Green (1995), Richardson and Green (1997)] , split/merge moves on the current clusters and creation/deletion of empty components - birth/death moves - will allow movements between different states. In particular, the idea summarises in the following. Being in state ψ, a move m is proposed and takes us to state ψ'. A vector of continuous random variables u, that are independent to ψ, are drawn and ψ' is set as a deterministic and invertible function of ψ and u. Move m is then accepted with probability

$$min\left\{1, \frac{p\left(\psi'|X\right)r_m\left(\psi'\right)}{p\left(\psi|X\right)r_m\left(\psi\right)q\left(u\right)}\left|\frac{\partial\psi'}{\partial\left(\psi,u\right)}\right|\right\},$$

where, $p(\psi|X)$ is the joint posterior density evaluated in state ψ, $r_m\left(\psi\right)$ is the probability of choosing move m when in state ψ and $q\left(u\right)$ is the density function of u.

Besides the creation/deletion of empty components, we have chosen a sampler that allows random split/merge moves. That means that we have designed the sampler such that it can randomly choose between both empty and non-empty components and then proceed to their splitting or merging. A closer look to the two types of moves, split/merge and birth/death, in the following sections will give us a better perspective of how the sampler is applied.

Split/Merge Moves

First in place is the choice between a split and a merge move. With probabilities b_G and d_G such that,

$$b_G = \begin{cases} 1, & \text{if} \quad G = 1 \\ 0.5, & \text{if} \quad G = 2, ..., G_{max} - 1 \\ 0, & \text{if} \quad G = G_{max} \end{cases}, \qquad d_G = \begin{cases} 0, & \text{if} \quad G = 1 \\ 0.5, & \text{if} \quad G = 2, ..., G_{max} - 1 \\ 1, & \text{if} \quad G = G_{max} \end{cases} \tag{2.4.2}$$

the sampler randomly chooses to either split a component into two or merge two components into one. The components to be split or merged can either be empty or non-empty.

In the case of a split move, a cluster ℓ, with weight w_ℓ, is randomly chosen and split into the clusters, say ℓ_1 and ℓ_2. Being in state ψ with parameters $\psi = (G, w, \gamma, y)$, the split move will now take us to state $\psi' = (G+1, w', \gamma, y')$. We can see that the variable selection vector γ remains the same, but we now have a new set of parameters for the clustering scheme. The weights of the newly formed components will be $w'_{\ell_1} = w_\ell u$ and $w'_{\ell_2} = w_\ell (1-u)$, where u is a random variable generated from a $Be\left(a_u, a_u\right)$. Following, the observations originally assigned to the selected cluster ℓ are randomly reallocated to the two new clusters, forming the updated allocation vector y'. Of course, the number of clusters is being increased by 1, $(G+1)$, while the number of the observations in the new clusters is n'_{ℓ_1} and n'_{ℓ_2} respectively, with $n'_{\ell_1} + n'_{\ell_2} = n_\ell$.

Considering the reverse move of randomly choosing two components, ℓ_1 and ℓ_2, and merging them into a single cluster, ℓ, with ℓ_1 and ℓ_2 being either empty or non-empty, we now move from state $\psi = (G+1, w, \gamma, y)$, to state $\psi' = (G, w', \gamma, y')$. All observations originally allocated to clusters ℓ_1 and ℓ_2, now belong in cluster ℓ, the weight of which is recalculated as $w'_\ell = w_{\ell_1} + w_{\ell_2}$, where w_{ℓ_1} and w_{ℓ_2} are the weights of ℓ_1,and ℓ_2 respectively. Evidently, $n'_\ell = n_{\ell_1} + n_{\ell_2}$, while this time the density of the random variable u is again a $Be\left(a_u, a_u\right)$ with $u = w_{\ell_1}/w'_\ell$.

The proposed split move is accepted with probability $\min\left(1, A\right)$, where A is the ratio :

$$A = \frac{p\left(\psi'|x\right) r_m\left(\psi'\right)}{p\left(\psi|x\right) r_m\left(\psi\right) q\left(u\right)} \left| \frac{\partial\left(w'_{\ell_1}, w'_{\ell_2}\right)}{\partial\left(w_\ell, u\right)} \right| .$$

In particular,

$$
\begin{aligned}
&p\left(\psi'|x\right) = f\left(G+1, w', \gamma, y'|X\right) = f\left(X, y'|w', \gamma, G+1\right) f\left(w'|G+1\right) f\left(G+1\right),\\
&p\left(\psi|x\right) = f\left(G, w, \gamma, y|X\right) = f\left(X, y|w, \gamma, G\right) f\left(w|G\right) f\left(G\right),\\
&r_m\left(\psi\right) = \frac{b_G p_{alloc}}{G}, \quad r_m\left(\psi'\right) = \frac{d_{G+1}}{G\left(G+1\right)},\\
&q\left(u\right) = \frac{1}{B\left(a_u, a_u\right)} u^{a_u-1}\left(1-u\right)^{a_u-1}, \quad \text{and} \quad \left|\frac{\partial\left(w'_{\ell_1}, w'_{\ell_2}\right)}{\partial\left(w_\ell, u\right)}\right| = w_\ell.
\end{aligned}
$$

With $p_{alloc} = u^{n'_{\ell 1}}\left(1-u\right)^{n'_{\ell 2}}$ being the probability that the particular allocation is made, substituting the above terms on the general form of the ratio A yields :

$$
\begin{aligned}
A = {} & \frac{f\left(X, y'|w', \gamma, G+1\right)}{f\left(X, y|w, \gamma, G\right)} \times \frac{f\left(w'|G+1\right)}{f\left(w|G\right)} \times \frac{f\left(G+1\right)}{f\left(G\right)}\\
& \times \frac{G d_{G+1}}{G\left(G+1\right) b_G p_{alloc}} \times \frac{B\left(a_u, a_u\right)}{u^{a_u-1}\left(1-u\right)^{a_u-1}} \times \frac{\left(G+1\right)!}{G!} \times w_\ell. \qquad (2.4.3)
\end{aligned}
$$

Finally, using the full conditionals in (2.3.5) and (2.3.6), and considering

$$
\frac{f\left(w'|G+1\right)}{f\left(w|G\right)} = \frac{w_{\ell_1}^{'\alpha_w-1} w_{\ell_2}^{'\alpha_w-1}}{w_\ell^{\alpha_w-1} B\left(\alpha_w, G\alpha_w\right)}, \quad \text{and} \quad \frac{f\left(G+1\right)}{f\left(G\right)} = 1,
$$

the acceptance ratios for both the cases of heterogeneous and homogeneous covariances are :

1. Homogeneity.

$$
\begin{aligned}
A = {} & \frac{\left(h_1 n'_{\ell_1}+1\right)^{\frac{-p_\gamma}{2}} \left(h_1 n'_{\ell_2}+1\right)^{\frac{-p_\gamma}{2}}}{\left(h_1 n_\ell+1\right)^{\frac{-p_\gamma}{2}}} \times \left[\frac{\left|Q^D + \sum_{k=1}^{G+1} S_k^{'D}\right|}{\left|Q^D + \sum_{k=1}^{G} S_k^D\right|}\right]^{\frac{-\left(n+\delta+p_\gamma-1\right)}{2}}\\
& \times \frac{B\left(a_u, a_u\right)}{B\left(\alpha_w, G\alpha_w\right)} \times u^{\alpha_w-a_u}\left(1-u\right)^{\alpha_w-a_u} \times w_\ell^{\alpha_w} \times \frac{d_{G+1}}{b_G}. \qquad (2.4.4)
\end{aligned}
$$

2. Heterogeneity.

$$\begin{aligned}A = {} & \left|Q^D\right|^{\frac{\delta+p\gamma-1}{2}} \times \frac{\left|Q^D+S_{l_1}^{\prime D}\right|^{\frac{-\left(n'_{l_1}+\delta+p\gamma-1\right)}{2}} \left|Q^D+S_{l_2}^{\prime D}\right|^{\frac{-\left(n'_{l_2}+\delta+p\gamma-1\right)}{2}}}{\left|Q^D+S_l^D\right|^{\frac{-\left(n_l+\delta+p\gamma-1\right)}{2}}} \\ & \times \frac{\left(h_1 n'_{\ell_1}+1\right)^{\frac{-p_\gamma}{2}} \left(h_1 n'_{\ell_2}+1\right)^{\frac{-p_\gamma}{2}}}{\left(h_1 n_\ell+1\right)^{\frac{-p_\gamma}{2}}} \\ & \times \prod_{j=1}^{p_\gamma} \frac{\Gamma\left(\frac{1}{2}\left(n_{l_1}+\delta+p_\gamma-j\right)\right)\Gamma\left(\frac{1}{2}\left(n_{l_2}+\delta+p_\gamma-j\right)\right)}{\Gamma\left(\frac{1}{2}\left(n_l+\delta+p_\gamma-j\right)\right)\Gamma\left(\frac{1}{2}\left(\delta+p_\gamma-j\right)\right)} \\ & \times \frac{B\left(a_u,a_u\right)}{B\left(\alpha_w,G\alpha_w\right)} \times u^{\alpha_w-a_u}\left(1-u\right)^{\alpha_w-a_u} \times w_\ell^{\alpha_w} \times \frac{d_{G+1}}{b_G}. \end{aligned} \tag{2.4.5}$$

The merge move, on the other hand, is accepted with probability min(1, A), with A being the inverse ratio of the one we obtained in the split move. More specifically we have :

$$\begin{aligned}A = {} & \frac{f\left(X,y'|w',\gamma,G\right)}{f\left(X,y|w,\gamma,G+1\right)} \times \frac{f\left(w'|G\right)}{f\left(w|G+1\right)} \times \frac{f\left(G\right)}{f\left(G+1\right)} \\ & \times \frac{G\left(G+1\right)b_G p_{alloc}}{G d_{G+1}} \times \frac{u^{a_u-1}\left(1-u\right)^{a_u-1}}{B\left(a_u,a_u\right)} \times \frac{G!}{\left(G+1\right)!} \times \left(w_{\ell_1}+w_{\ell_2}\right)^{-1}, \end{aligned} \tag{2.4.6}$$

which can be further simplified for the cases of homogeneity and heterogeneity by taking the inverse ratios of (2.4.4) and (2.4.5) respectively.

Birth/Death Moves

In the final part of the parameters' sampling we have the second set of moves, the birth/death moves. Like in the split/merge process, a birth or death move is randomly selected with probabilities b_{G_0} and d_{G_0} :

$$b_{G_0} = \begin{cases} 1, & \text{if} \quad G_0 = 0 \\ 0.5, & \text{if} \quad G_0 = 2, ..., G_{max} - 2 \\ 0, & \text{if} \quad G_0 = G_{max} - 1 \end{cases}, \quad d_{G_0} = \begin{cases} 0, & \text{if} \quad G_0 = 0 \\ 0.5, & \text{if} \quad G_0 = 2, ..., G_{max} - 2 \\ 1, & \text{if} \quad G_0 = G_{max} - 1 \end{cases} \tag{2.4.7}$$

where G_0 is the number of empty components before we proceed to the birth/death move.

Under a birth move, a new empty component is generated with weight w'_{G+1} drawn from a $Be\,(1, G)$. The existing weights w_k, then need to be rescaled to ensure summation to unity. The new weights will now be $w'_k = w_k \left(1 - w'_{G+1}\right)$, for $k = 1, \ldots, G$. In terms of movement between states, starting from $\psi = (G, w, \gamma, y)$, the birth of the new component takes us into state $\psi' = (G + 1, w', \gamma, y)$, where we note that the allocation vector y now remains the same since no observations need to be reallocated. The move is accepted with probability $\min(1, A)$, where A stands for :

$$A = \frac{p\,(\psi'|x)\, r_m\,(\psi')}{p\,(\psi|x)\, r_m\,(\psi)\, q\left(w'_{G+1}\right)} \left|\frac{\partial(w')}{\partial\left(w, w'_{G+1}\right)}\right|.$$

This time the terms of the ratio A are :

$$p\,(\psi'|x) = f\,(G + 1, w', \gamma, y|X) = f\,(X, y|w', \gamma, G + 1)\, f\,(w'|G + 1)\, f\,(G + 1)\,,$$

$$p\,(\psi|x) = f\,(G, w, \gamma, y|X) = f\,(X, y|w, \gamma, G)\, f\,(w|G)\, f\,(G)\,,$$

$$r_m\,(\psi) = b_{G_0}, \quad r_m\,(\psi') = \frac{d_{G_0+1}}{G_0 + 1},$$

$$q\left(w'_{G+1}\right) = G\left(1 - w'_{G+1}\right)^{G-1}, \quad \left|\frac{\partial\,(w')}{\partial\left(w, w'_{G+1}\right)}\right| = \left(1 - w'_{G+1}\right)^{G-1},$$

which yield :

$$\begin{aligned} A = & \frac{f\left(X,y|w',\gamma,G+1\right)}{f\left(X,y|w,\gamma,G\right)} \times \frac{f\left(w'|G+1\right)}{f\left(w|G\right)} \times \frac{f\left(G+1\right)}{f\left(G\right)} \\ & \times \frac{d_{G_0+1}}{(G_0+1)b_{G_0}} \times \frac{1}{G\left(1-w'_{G+1}\right)^{G-1}} \times \frac{(G+1)!}{G!} \times \left(1-w'_{G+1}\right)^{G-1}. \end{aligned} \tag{2.4.8}$$

With further calculations we obtained the final form of the ratio A for the birth move, which applies to both the cases of homogeneous and heterogeneous covariances. That is:

$$A = \left(1-w'_{G+1}\right)^{n} \times \frac{w_{G+1}^{\prime(\alpha_w-1)}\left(1-w'_{G+1}\right)^{G(\alpha_w-1)}}{B\left(\alpha_w,G\alpha_w\right)} \times \frac{d_{G_0+1}}{\left(G_0+1\right)b_{G_0}} \times \frac{G+1}{G}. \tag{2.4.9}$$

Finally, in the case of a death move, an empty component, say w_{G+1} is randomly selected and deleted. Again, we need to rescale the remaining weights, but this time $w'_k = w_k/(1-w_{G+1})$, for $k=1,\ldots,G$. The move, which takes us from state $\psi = (G+1,w,\gamma,y)$ to state $\psi' = (G,w',\gamma,y)$, is accepted with probability $\min\left(1,A\right)$, where A is the inverse ratio of the birth move, i.e :

$$A = \left(1-w_{G+1}\right)^{-n} \times \frac{B\left(\alpha_w,G\alpha_w\right)}{w_{G+1}^{(\alpha_w-1)}\left(1-w_{G+1}\right)^{G(\alpha_w-1)}} \times \frac{\left(G_0+1\right)b_{G_0}}{d_{G_0+1}} \times \frac{G}{G+1}. \tag{2.4.10}$$

At this point we should refer to a problem usually met on the application of mixture models within the Bayesian framework, the so-called label switching problem. Redner and Walker (1984) described the problem that arises under permutations of the mixture components due to the invariance of the likelihood. Within the Bayesian framework the latter could lead to posterior distributions characterised by symmetry and multimodality. Being interested in estimating parameters summarizing over the posterior distribution, we understand that such a phenomenon can complicate inference on the parameters of interest producing inappropriate estimations. A common solution to this problem is setting inequality constraints on the parametric space (Diebolt and Robert, 1994), while, others have developed relabelling algorithms [(Stephens, 2000), (Celeux, 1998)] or label invariant loss functions (Celeux et al., 2000). Jasra et al. (2005) review different approaches on addressing the problem of label switching.

Of course, label switching is a phenomenon we encounter when clustering using mixture models, as labels permute along our Markov chains. However, as far as the methodology presented here is concerned, we have integrated the component means and covariances out of the posterior and we are interested only in the estimation of the allocation vector y (together with the variable selection vector γ). To estimate the cluster structure of our observations, however, we do not compute the marginal posterior probabilities of an observation i being allocated to cluster k, but rather, we examine the probabilities of two observations being allocated to the same cluster, $P(y_i = y_j)$, which is invariant to permutations. Using these probabilities we then form clustering maps that indicate the cluster formulation of our sample and from which further inference on whether our model has converged to the expected cluster structure can be drawn.

2.5 Simulation Study

It is now time to examine the performance of the model on a set of simulated points. We chose the case of heterogeneous covariance matrices and sampled a data set of 15 observations. The total number of variables is $50, 20$ of which have been chosen to be discriminating. The sample is designed as coming from four groups of different means and covariances for the 20 discriminating variables, and as a set of observations that favour a single multivariate distribution for the remaining 30 noisy variables. A standard multivariate normal $N(0, I_{30})$ was used for the generation of the noisy variables, while the discriminating ones were drawn from four different multivariate normals. More specifically, the first four observations have been assigned to the first group following a Normal distribution with $\mu_1 = 5$ and $\sigma_1^2 = 1.5$, the following three to the second one with $\mu_2 = 2$ and $\sigma_2^2 = 0.1$, the third group is consisted by the next six with $\mu_3 = -3$ and $\sigma_3^2 = 0.5$ and the last one by the last two observations with $\mu_4 = -6$ and $\sigma_4^2 = 2$, i.e.

$$x_{ij} \sim I_{1\leq i\leq 4} N\left(5, 1.5\right) + I_{5\leq i\leq 7} N\left(2, 0.1\right) + I_{8\leq i\leq 13} N\left(-3, 0.5\right) + I_{14\leq i\leq 15} N\left(-6, 2\right),$$
$$i = 1, \ldots, 15, j = 1, \ldots, 20 \quad .$$

A similar design is also used in Tadesse et al. (2005).

Having analysed the structure of the data set, we now come to the hyperparameters of the model and their values. The first hyperparameters to be decided is the set referring to the Beta - Binomial prior on the total number of discriminating variables, p_γ. We set the a priori expected number of discriminating variables p_{prior} equal to 10, while for the parameters of $Be\left(a, b\right)$, we chose a, b such that $a + b = 2$ to ensure a vague prior. We allow $G_{max} = 15$ for the discrete uniform on the number of components and a shape parameter of $\delta = 3$ for the Inverse-Wishart distributions on the covariance matrices.

We continue with the values for the remaining hyperparameters of the Inverse-Wishart distributions, which are the scalars h_1, h_0 and the percentages within the hyperparameters κ_1, κ_0. Recall, κ_1 is the $1\% - 10\%$ of the upper decile of the $n - 1$ non-zero eigenvalues of $cov\left(X\right)$ and κ_0 is, respectively, the percentage of their lower decile. Let us, indicate the percentages in κ_1 and κ_0 with the scalars c_1 and c_0 respectively. Several different values have been tried for this set of hyperparameters and in particular, $h_1 = h_0 = 10, 100, 1000$, with $c_1 = c_0 = 1\%, 3\%, 9\%$. Table 1 indicates the value of the diagonal elements of the Q^D, Q^{ND}

scale matrices, for the three cases of $c_1 = c_0 = 1\%, 3\%$ and 9%.

	$c_1 = c_0$		
	0.01	0.03	0.09
Q_{ii}^{D}	2.4583	0.8194	0.2731
Q_{ii}^{ND}	96.4221	32.1407	10.7136

Table 1: Diagonal elements of the Q^D, Q^{ND} scale matrices of the simulated data set for the three cases of $c_1 = c_0 = 0.01, 0.03, 0.09$.

We should point out here, that even though we tested the algorithm for different combinations of the pairs (h_1, h_0) and (c_1, c_0), we find that there is no need for setting $h_1 \neq h_0$ or $c_1 \neq c_0$. But, allow us to conclude this section with the choice of α_w and a_u, and we shall return to the settings of these critical parameters with further comments in the forthcoming paragraphs. The values we tried for the parameter of the prior on the component weights α_w are $(1, 6)$, while for the α_u of the $Be\left(a_u, a_u\right)$ we used in the split/merge moves we tried values such as $(2, 6)$.

We ran the algorithm for 60000 iterations, with a burn-in of 40000 and produced graphs to illustrate the selection of variables, as well as the cluster structure of the data. On a two quad core server with 2.53Ghz CPUs, the code needed around 3 hours. As far as the variable selection task is concerned, we use histograms that indicate the total number of discriminating variables, while we plot the marginal posterior probabilities for the $p = 50$ variables of the simulated data set to show the probability of each variable being chosen to be included in the model. For the current simulated example and under the assumption of a correctly performing model, we expect histograms that indicate the selection of $p_\gamma = 20$ discriminating variables, with the first 20 variables having high marginal posterior probabilities (close to 1), while low posterior probability (close to 0) are assigned to each of the remaining 30 variables. Finally, for all $n = 15$ observations, we calculate the probabilities of two observations being allocated in the same group, i.e. $P(y_i = y_j), i, j = 1, \ldots, 15$, and colour them with a scale for which black indicates low probability and white high. A clustering map is therefore produced, where observations allocated in the same cluster form white patches. Considering the four groups described earlier, we expect a map of four white patches on the diagonal, each of which must be formed by $4, 3, 6$ and 2 observations respectively.

Getting back to the settings of the runs, we start the algorithm assuming the starting point of all variables being included in the model, while each observation forms a single group. However, we noticed that the algorithm had some difficulties in uncovering the cluster structure of the data and recovering the 20 discriminating variables. We, therefore, experimented with setting the starting points of γ and y equal to the vectors according to which the simulated data set was originally designed, i.e. 20 discriminating variables for γ with the observations coming from 4 groups (y). Looking at Figures A.0.10, A.0.11 and A.0.12 in Appendix A, one can draw the conclusion that the different starting values did not particularly improve the results.

Regarding how the selected values of the hyperparameters affect the algorithm, some very interesting remarks can be made at this point. First of all, we can note that the different values on α_w and a_u do not alter the resulting inference on either the clustering or the variable selection tasks (Figures A.0.1 - A.0.9 in Appendix A). However, the choice of h_1, h_0 and c_1, c_0 appears to be of great importance. Figures 2.5.1, 2.5.2 and 2.5.3 illustrate the impact of the different values of these hyperparameters in the recovering of the clusters and the discriminating variables. In Figure 2.5.1 we can see the histograms for the number of variables included in the model (p_γ), while Figure 2.5.2 indicates the posterior inclusion probabilities for each of the 50 variables. Finally, Figure 2.5.3 shows the corresponding favoured cluster structure of the observations, for the different combinations of h_1, h_0, c_1, and c_0. Since, as stated earlier, we considered $h_1 = h_0$ and $c_1 = c_0$, for simplicity reasons, in the graphs below we used the symbols h and c to denote the values referring to h_1, h_0 and c_1, c_0 respectively. The starting values used for this set of graphs are : all the variables are included in the model and every observation forms a single cluster, while, we have chosen $\alpha_w = 1$ and $a_u = 2$. Similar figures with a different set of starting values can be found in Appendix A(Figures A.0.10 - A.0.12).

We can see that the algorithm recovers the correct set of discriminating variables as well as uncovers the correct clustering scheme, only when c is set equal to 0.09, and in the exceptional case of $c = 0.03$ with a h equal to 1000. Evidently, the choice of h (or h_1 and h_0) and c (or c_1 and c_0) is very crucial and extra care is needed to be taken from the researcher as there is no rule of thumb that one can follow for the selection of those values.

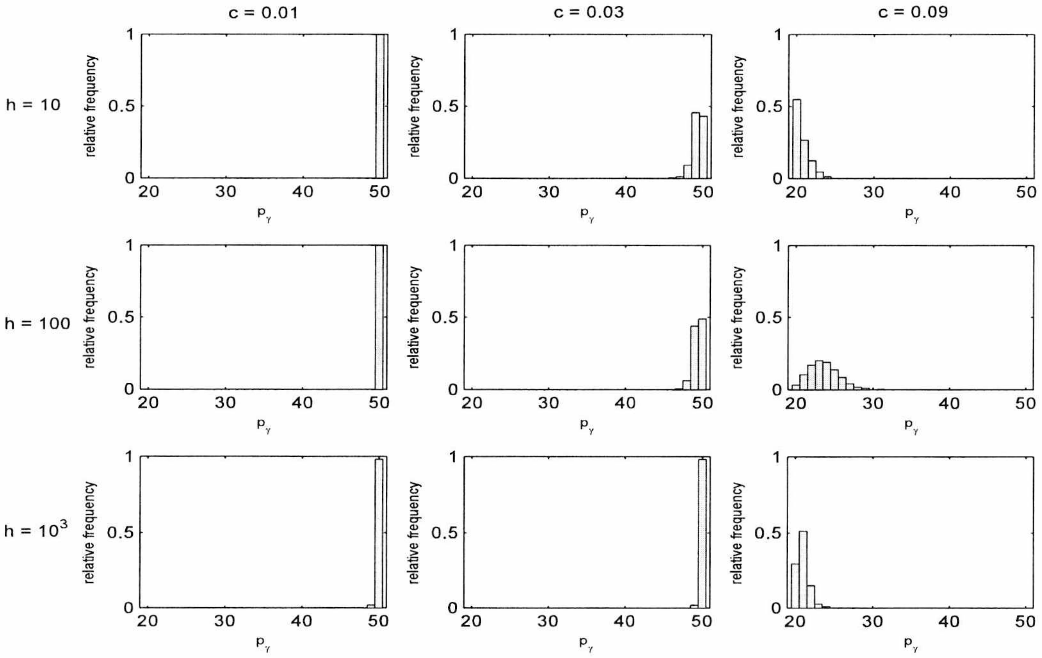

Figure 2.5.1: Simulated data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices.

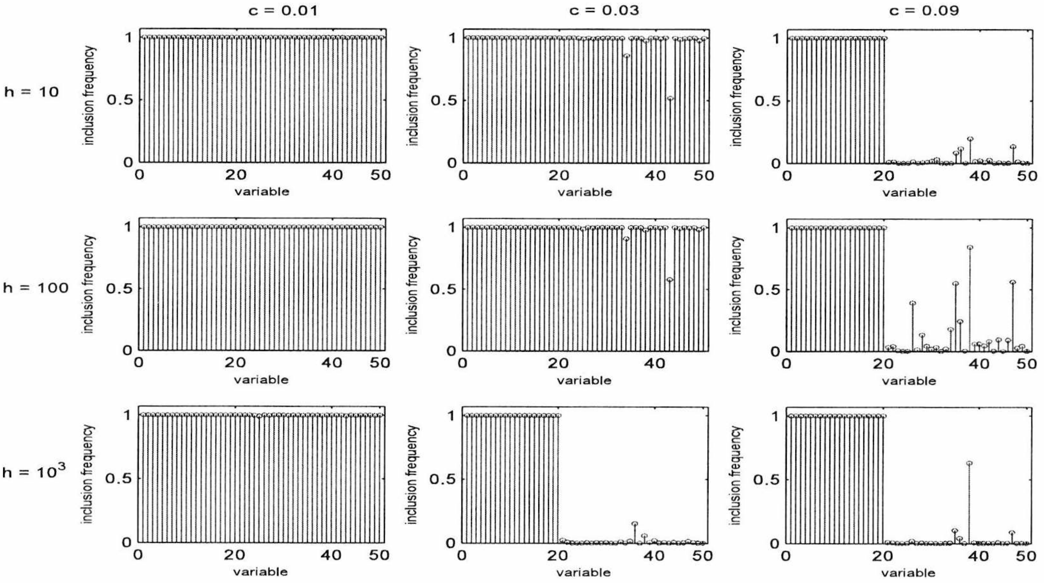

Figure 2.5.2: Simulated data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices.

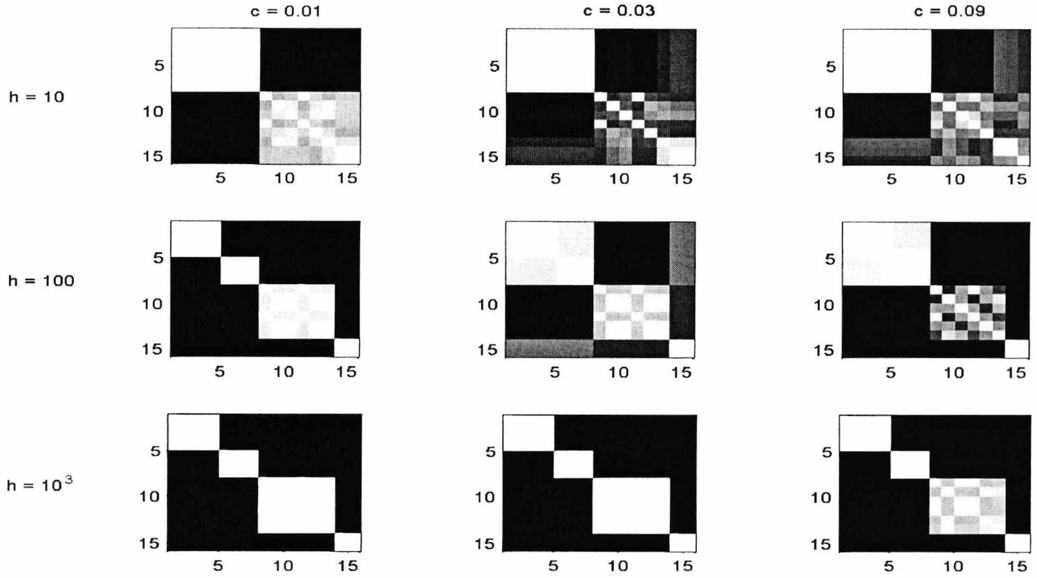

Figure 2.5.3: Simulated data: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices.

On the efficiency of the RJMCMC sampler and for $\alpha_u = 2$, the average acceptance rates for the split, merge, birth and death moves were $0.3771, 0.4845, 0.3562$, and 0.7825 respectively. Although one would think the choice of the parameter α_u as important in determining the mixing of the above moves, the results do not indicate so. With $\alpha_u = 6$ the average acceptance rates are similar to the ones above. More specifically, we have the rates $0.2893, 0.3895, 0.3501$, and 0.7798 for the split, merge, birth and death moves respectively.

2.6 Preprocessing on the Simulated Data

On account of the conclusions drawn about the importance of the hyperparameters h_1, h_0 and c_1, c_0, and given the lack of a useful tool that could indicate a good set of values, a two-stage analysis, in which stage 0 consists of preliminaries to specify these hyperparameters, could be of great practical use.

The idea commences with the preprocessing stage (stage 0), in which a Principal Component Analysis is implemented to reduce the dimensionality of the problem. The principal components admitted to the analysis that follows, say m^*, can be chosen according to either the variance explained by the first m^* principal components, or based on an "elbow" that can be observed when plotting these variances.

After m^* has been chosen, stage 0 continues by performing a K-means analysis on the selected components. Assessing the results of the K-means procedure, implications can be made about the values that could be used for the hyperparameters of interest (h_1, h_0, c_1, c_0). An idea is to exploit the grouping suggested by K-means and calculate the new within and between cross products matrices for the set of the original variables. Using the cross products matrices, a new diagonal matrix Q can be formed, from which the scale matrices Q^D and Q^{ND} of the Inverse-Wishart priors can be extracted.

More specifically, for the clusters suggested by K-means, we calculate the covariance matrices, say D_k is the covariance matrix of the k^{th} group, and then calculate their pooled mean, i.e.

$$S_{pool} = \frac{\sum_{k=1}^{G} (n_k - 1) D_k}{\sum_{k=1}^{G} (n_k - 1)},$$

where, n_k is the number of observations allocated in the k^{th} group. The diagonal elements of matrix S_{pool} are then used to form a new matrix Q. Taking the diagonal elements of Q that refer to the discriminating variables, we build the diagonal scale matrix Q^D of the Inverse-Wishart prior assigned to Σ_k^D. Similarly, a scale matrix Q^{ND} is built for the Inverse-Wishart on the covariance matrix of the non-discriminating variables. Matrices S_{pool} and Q are of dimension $p \times p$, while diagonal matrices Q^D and Q^{ND} have dimensions p_γ and $p - p_\gamma$ respectively (where p is the total number of variables).

As far as h_1 and h_0 are concerned, values that illustrate the variability of the groups would be much preferred. Once again, for the groups suggested by the K-means clustering and for the m^* principal components, we calculate the between groups covariance matrix, say $S_{between}$. We then estimate h_1 as the mean :

$$\hat{h_1} = \frac{1}{m^*} \sum_{i=1}^{m^*} \frac{(S_{between})_{ii}}{(S_{pool})_{ii}}. \tag{2.6.1}$$

Finally, we set $h_0 = h_1$.

Having completed the preliminary analysis of stage 0, we then make use of the estimators of the crucial hyperparameters and perform the full MCMC procedure described in section 2.4. Stage 1 will consist of the preprocessing and will determine the estimation of the parameters of interest γ, y, w and G.

We now apply the preprocessing procedure on the simulated data set. Starting with the Principal Component Analysis using the covariance matrix, the variance of the data explained by the resulting components can be seen in Table 2; for spatial reasons, the table contains only the first 5 principal components and their corresponding percentage of the variance explained.

	Principal Component				
	1	2	3	4	5
Variance Explained	88.35 %	90.59 %	92.21 %	93.74 %	95%

Table 2: Variance explained by the first 5 principal components of the simulated data set.

We can see that the biggest chunk of the variance of the data is explained by the first principal component (88%).We thus decided to proceed the analysis keeping this first component only. Similar decision can be drawn by looking at the scree plot in Figure 2.6.1, where one can see the "elbow" breaking the line in the second principal component.

Carrying on with the K-means clustering on the first principal component with a prespecified number of groups of four, the observations were assigned in groups resembling the original cluster structure of the data. Implementing the suggested grouping we followed the scheme for specifying Q^D, Q^{ND}, h_1 and h_0. The estimated value of h_1 and in continuity of h_0 was 983.26, while the estimated diagonal elements of Q^D and Q^{ND} can be seen in Figure 2.6.2. Making use of the resulting estimates, we then ran the full MCMC on the original data set.

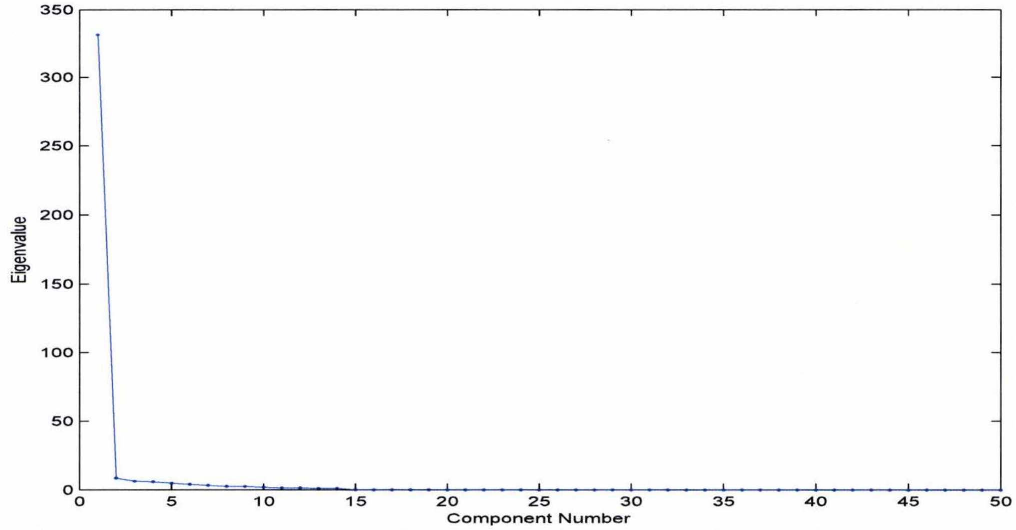

Figure 2.6.1: Scree plot of the principal components for the simulated data set.

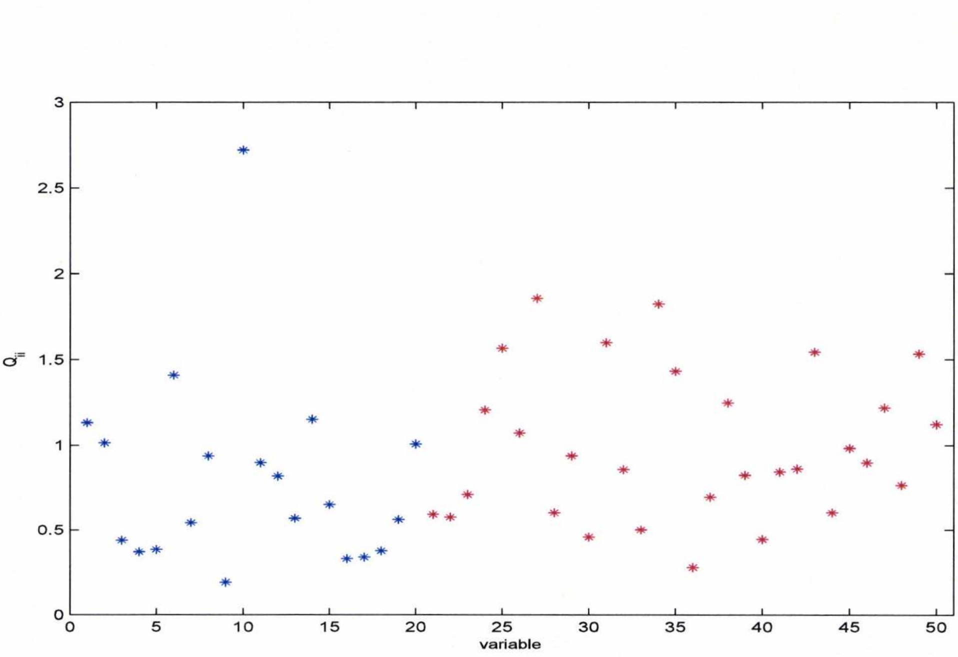

Figure 2.6.2: Estimated diagonal elements for the matrices Q^D - indicated with blue colour - and Q^{ND} - indicated with red colour.

Considering the starting values of all variables being included in the model and each observation assigned to a single group, we run the algorithm again for 60000 iterations with a burn-in of 40000. Since, as we saw in the application of the previous section, the parameters α_w and a_u do not affect the resulting extraction of discriminating variables or clustering, we considered the values $\alpha_w = 1$ and $a_u = 2$. For a better insight of the problem, additionally to the estimated values for h_1, h_0, keeping Q^D, Q^{ND} as estimated in stage 0, we examined the results for $h_1 = h_0 = 10, 100$, and 1000. Of course values of $h_1 = h_0 = 983.26$ and 1000 give similar results for both the clustering and the variable selection task. Figures 2.6.3 and 2.6.4, show that none of the cases we examined could extract the 20 discriminating variables; for all cases we have the whole set of 50 variables included in the model. However, the four clusters could be identified for all cases except when $h_1 = h_0 = 10$ (Figure 2.6.5).

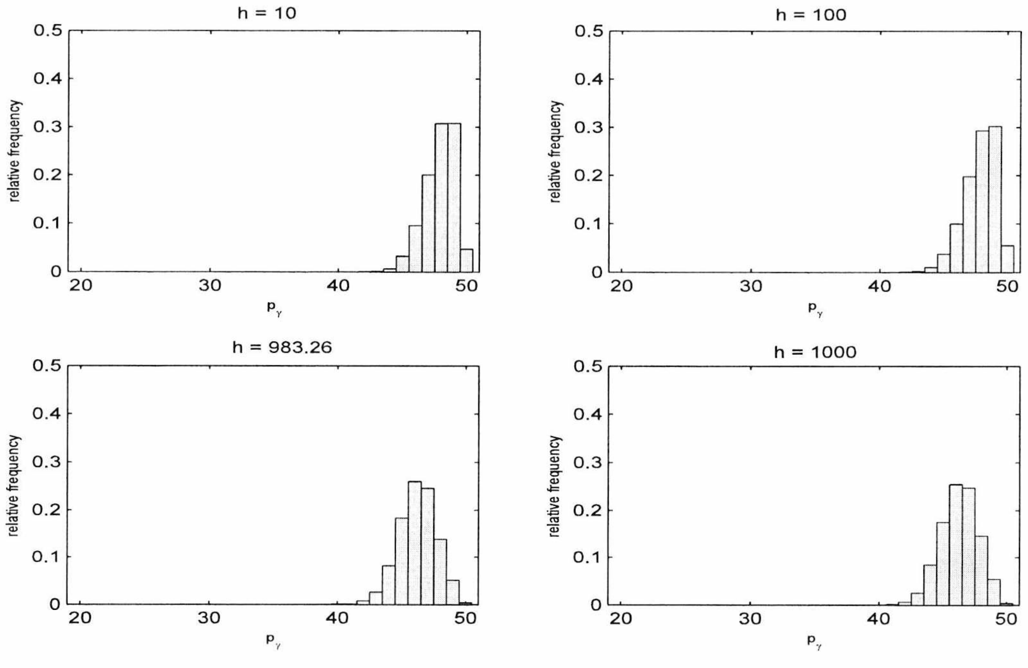

Figure 2.6.3: Two-Stage Preprocess: Histograms of the total number of discriminating variables, p_γ, for the simulated data, with $h_1 = h_0 = 10, 100, 983.26, 1000$, assuming unequal covariance matrices.

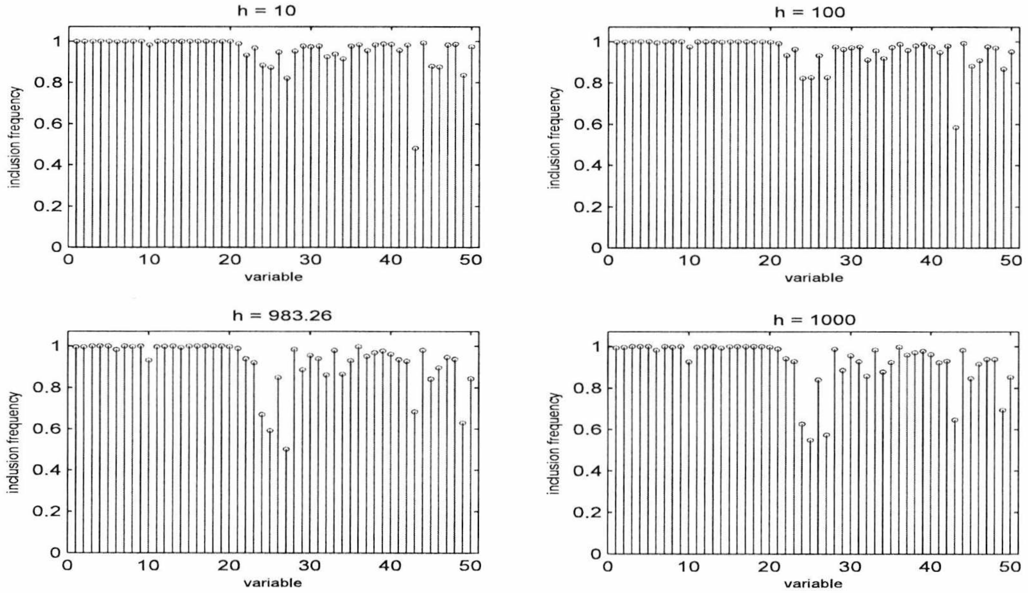

Figure 2.6.4: Two-Stage Preprocess: Marginal Posterior Probabilities of the variables included in the model for the simulated data, with $h_1 = h_0 = 10, 100, 983.26, 1000$, assuming unequal covariance matrices.

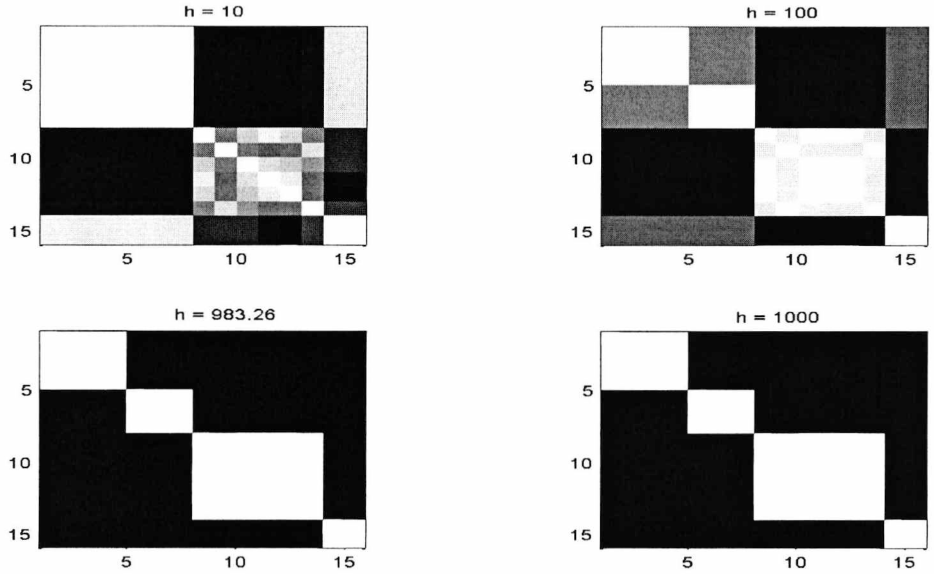

Figure 2.6.5: Two-Stage Preprocess: Maps of the cluster allocations of the $n = 15$ observations of the simulated data, with $h_1 = h_0 = 10, 100, 983.26, 1000$, assuming unequal covariance matrices.

Although the cluster structure of the observations could be recovered for the estimated value of h_1 and h_0, the unresolved problem of the selection of the important variables indicates the of crucial importance choice of the scale matrices Q^D and Q^{ND}.

2.7 Split/Merge moves using the SAMS sampler

In our attempts so far, the resulting figures demonstrate that the algorithm has a fair difficulty in uncovering the cluster structure of the simulated data set, although it is a trivial example, as the observations have been sampled such that the clusters are well separated and thus easily identifiable. In particular, we often observe the first two clusters being merged together, while the third is being commonly split into further smaller components. Consequently, one might presume that there is a complication with the split/merge moves of the Reversible Jump process. In need of a computationally more efficient procedure of suggesting split and merge moves under the Reversible Jump methodology, we considered the use of the split-merge sampler (Sequentially-Allocated Merge-Split Sampler) proposed by Dahl (2005). Proposing split moves quickly, needing no further sweetening, as Dahl (2005) puts it, the SAMS sampler borrows ideas from sequential importance sampling and draws samples from the correct stationary distribution for Bayesian nonparametric models. For both conjugate and nonconjugate Dirichlet process mixture (DPM) models, SAMS proposes splits by sequentially allocating observations using probabilities that condition on previously allocated data.

Keeping the MCMC steps for the sampling of the parameters (γ, y, w) and yet considering the birth and death of empty components, we chose to suggest splits similarly to the conjugate SAMS. We understand that the only change induced in the model presented so far, is the acceptance probabilities of the split and merge moves in (2.4.3) and (2.4.6) respectively. We describe the new moves and recalculate the acceptance probabilities in the following paragraphs.

The sampler begins by uniformly selecting two observations, say i and j. If i and j belong to the same group, the sampler proceeds to a split move, whereas, if i, j come from two different groups, a merge move is proposed. Evidently, a generation or deletion of a component will effectively change the dimension of the component weights, which need to be updated. We therefore considered the conjugate case of the SAMS sampler.

Let us first examine the case of a split move. Letting i, j come from the same group S, a split move begins by allocating the two observations in two separate groups S^i, S^j. Each of the remaining observations in group S, say t, is then being allocated to S^i with probability :

$$Pr\left(y_t = S^i | y^{S^i}, y^{S^j}\right) = \frac{f\left(y_t = S^i | X, y_{(-t)}, \gamma, w, G\right)}{f\left(y_t = S^i | X, y_{(-t)}, \gamma, w, G\right) + f\left(y_t = S^j | X, y_{(-t)}, \gamma, w, G\right)}. \tag{2.7.1}$$

As we have stated before, $y_{(-t)}$ represents the cluster allocations of all the observations except for the observation t, while, we introduce y^{S^i} and y^{S^j} to indicate the observations assigned to group S^i and S^j respectively. We should also note that $f\left(y_t = S^i | X, y_{(-t)}, \gamma, w, G\right)$ is the posterior probability (2.4.1). For an observation t allocated in group S, we can generally write :

$$f\left(y_t | y_{(-t)}\right) = \frac{f\left(y^S, y_t\right) \prod_{\substack{k=1 \\ k \neq S}}^{G} f\left(y^k\right)}{\prod_{k=1}^{G} f\left(y^k\right)} = \frac{f\left(y^S, y_t\right)}{f\left(y^S\right)}.$$

Therefore, the terms of (2.7.1) can be rewritten giving the form :

$$Pr\left(y_t = S^i | y^{S^i}, y^{S^j}\right) = \frac{\frac{f\left(X, y^{S^i}, y_t | G, w, \gamma\right)}{f\left(X, y^{S^i} | G, w, \gamma\right)}}{\frac{f\left(X, y^{S^i}, y_t | G, w, \gamma\right)}{f\left(X, y^{S^i} | G, w, \gamma\right)} + \frac{f\left(X, y^{S^j}, y_t | G, w, \gamma\right)}{f\left(X, y^{S^j} | G, w, \gamma\right)}}, \tag{2.7.2}$$

the analytic form of which - after the necessary calculations - is given by (A.3) in Appendix A. Finally, t is being assigned to S^j with probability

$$Pr\left(y_t = S^t | y^{S^i}, y^{S^j}\right) = 1 - Pr\left(y_t = S^i | y^{S^i}, y^{S^j}\right). \tag{2.7.3}$$

After all the observations initially allocated in the group to be split S, have been reallocated to either S^i or S^j, we have two newly formed components, the weights of which, w_{S^i} and w_{S^j} need to be updated. Dahl (2005) suggests that the new values of the parameters related to group S^i are proposed by either sampling from the centering distribution F_0 of the DPM model, or by using a random walk. However, we choose to comply with the idea of updating the new component weights via a transformation of the weight of the component to be split (w_S). With a random variable $u \sim Be\left(a_u, a_u\right)$, we generate $w'_{S^i} = w_S u$ and $w'_{S^j} = w_S\left(1 - u\right)$.

On the Metropolis-Hastings acceptance ratio of move m,

$$a = \min\left\{1, \frac{p\left(\psi' | x\right) r\left(\psi | \psi'\right)}{p\left(\psi | x\right) r\left(\psi' | \psi\right)}\right\}, \tag{2.7.4}$$

$p\left(\psi | x\right)$ is the joint posterior distribution when in state ψ and $r\left(\psi | \psi'\right)$ the probability of proposing the move from state ψ' to state ψ. With $\psi = \left(G, w, \gamma, y\right)$ and $\psi' = \left(G + 1, w, \gamma, y\right)$, where $w = \left(w_1, \ldots, w_S, \ldots, w_G\right)$ and $w' = \left(w'_1, \ldots, w'_{S^i}, w'_{S^j}, \ldots, w'_{G+1}\right)$, $r\left(\psi' | \psi\right)$ is the product of the allocation probabilities (2.7.2), (2.7.3), multiplied by the proposal density that takes us to the updated parameters $\left(w'_{S^i}, w'_{S^j}\right)$.

Since the two split components could only be merged in one way, the reverse proposal probability $r(\psi|\psi')$ is always 1. Taking into account the density function of u, $q(u)$, the Jacobian arising from the transformation of the weights, the allocation probabilities, and allowing the acceptance probability in (2.7.4) to be written as $a = \min(1, A)$, ratio A becomes :

$$A = \frac{p(\psi'|x)}{p(\psi|x)\, q(u) \prod_{t \in S^i} Pr\left(y_t' = S^i | y'^{S^i}, y'^{S^j}\right) \prod_{t \in S^j} Pr\left(y_t' = S^j | y'^{S^i}, y'^{S^j}\right)} \left| \frac{\partial\left(w'_{S^i}, w'_{S^j}\right)}{\partial\left(w_S, u\right)} \right| . \tag{2.7.5}$$

Going one step further and substituting the joint posteriors $p(\psi'|x)$, $p(\psi'|x)$, the Jacobian and the proposal of u, it yields :

$$\begin{aligned} A &= \frac{f(X, y'|w', \gamma, G+1)}{f(X, y|w, \gamma, G)} \times \frac{f(w'|G+1)}{f(w|G)} \times \frac{f(G+1)}{f(G)} \\ &\times \frac{1}{\prod_{t \in S^i} Pr\left(y_t' = S^i | y'^{S^i}, y'^{S^j}\right) \prod_{t \in S^j} Pr\left(y_t' = S^j | y'^{S^i}, y'^{S^j}\right)} \\ &\times \frac{1}{B(a_u, a_u)} \frac{1}{u^{a_u - 1}(1-u)^{a_u - 1}} w_S . \end{aligned} \tag{2.7.6}$$

Formulas (A.4) and (A.5) in Appendix A give ratio A for the cases of Homogeneous and Heterogeneous covariance matrices respectively.

We have, so far, seen the split move of the SAMS sampler. However, when the selected observations i, j, belong to different groups, say S^i and S^j respectively, a merge move is proposed. This time, the observations of S^i and S^j are being reallocated in a single group S, with weight $w'_S = w_{S^i} + w_{S^j}$; note here that Dahl (2005) suggests the model parameter of the new component S equal to the model parameter of group S^j. The merging of the components can be made only in one way and thus the proposal density $r(\psi'|\psi)$, with $\psi = (G+1, w, \gamma, y)$ and

$\psi' = (G, w', \gamma, y')$, is equal to 1. For the reverse probability though, we need to consider the merged group of state ψ', as a component to be split in two further components. Beware that the proposed partition is needed to resemble the partition of the observations originally set in S^i and S^j, while, the random variable u is now $u = w_{S^i}/w'_S$. The allocation probabilities in (2.7.2) and (2.7.3), will be resulting to the desirable partition and their product, multiplied by the proposal density that presumably gives the weights w_{S^i}, w_{S^j}, will now yield the reverse proposal probability $r(\psi|\psi')$. In other words, the merge move is accepted with probability $a = \min(1, A)$, where A, as results from the Metropolis - Hastings ratio in (2.7.4), is :

$$A = \frac{p(\psi'|x)\, q(u) \prod_{t\in S^i} Pr\left(y_t = S^i|y^{S^i}, y^{S^j}\right) \prod_{t\in S^j} Pr\left(y_t = S^j|y^{S^i}, y^{S^j}\right)}{p(\psi|x)} \left|\frac{\partial(w'_S, u)}{\partial(w_{S^i}, w_{S^j})}\right|, \tag{2.7.7}$$

which, further becomes :

$$\begin{aligned} A = {} & \frac{f(X, y'|w', \gamma, G-1)}{f(X, y|w, \gamma, G)} \times \frac{f(w'|G-1)}{f(w|G)} \times \frac{f(G-1)}{f(G)} \\ & \times \prod_{t\in S^i} Pr\left(y_t = S^i|y^{S^i}, y^{S^j}\right) \prod_{t\in S^j} Pr\left(y_t = S^t|y^{S^i}, y^{S^j}\right) \\ & \times B(a_u, a_u)\, u^{a_u-1} (1-u)^{a_u-1}\, w_S'^{-1}. \end{aligned} \tag{2.7.8}$$

Taking a closer look in equation (2.7.8), we notice that the ratio A of the merge move is the reverse of that of the split move. Therefore, the reverse of the formulas (A.4) and (A.5) in Appendix A will give us the resulting ratios of the merge move for the case of Homogeneous and Heterogeneous covariance matrices respectively.

We conclude this section with the application of the altered algorithm on the simulated data set of section 2.5. With $\alpha_w = 1$, $a_u = 2$ and 60000 iterations with a burn-in of 40000, we estimated the number of discriminating variables and the cluster allocations of the 15 observations starting with all the variables included in the model and each observation assigned to a single group. The resulting histograms, posterior probability plots and clustering maps, estimated for the cases of $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, in Figures 2.7.1, 2.7.2 and 2.7.3 respectively, demonstrate the difficulty met on the identification of the important variables, as well as the masked cluster structure. We can see that even for the case of $h = 1000$, for which the four groups could be recovered on the application of the same data set in 2.5, there is only the unique case of $c = 0.09$ that succeeds in the recovering of the cluster structure of the data.

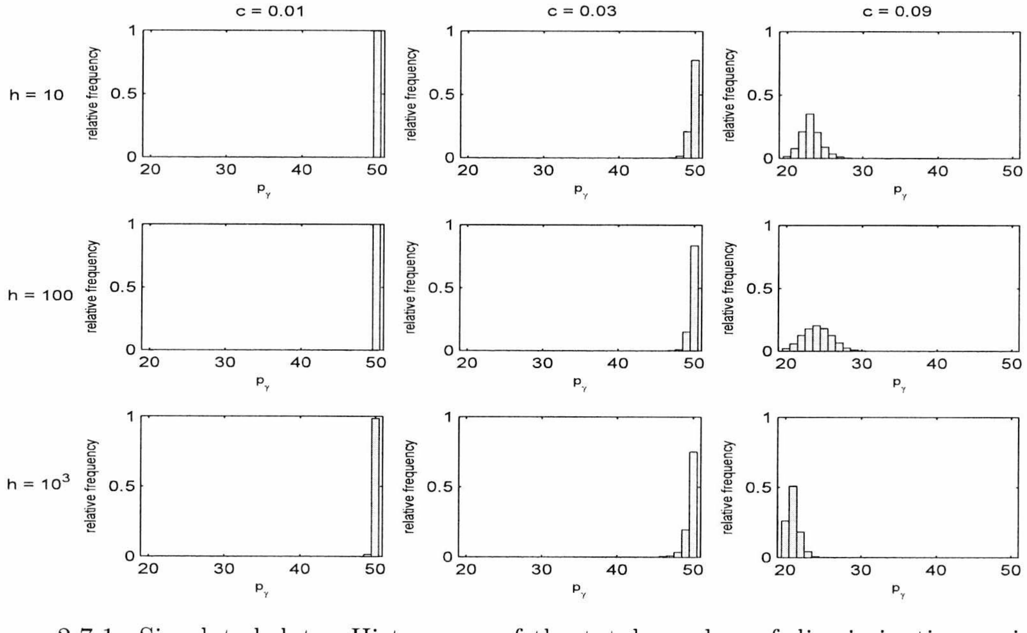

Figure 2.7.1: Simulated data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

This time, the acceptance rates of the sampler for the split and merge moves are low. In particular we have split and merge moves with average acceptance rates 0.0158 and 0.1187 respectively. Also, for the birth and death moves we have rates 0.3264 and 0.8159, while the CPU time used for the above runs was 3 hours.

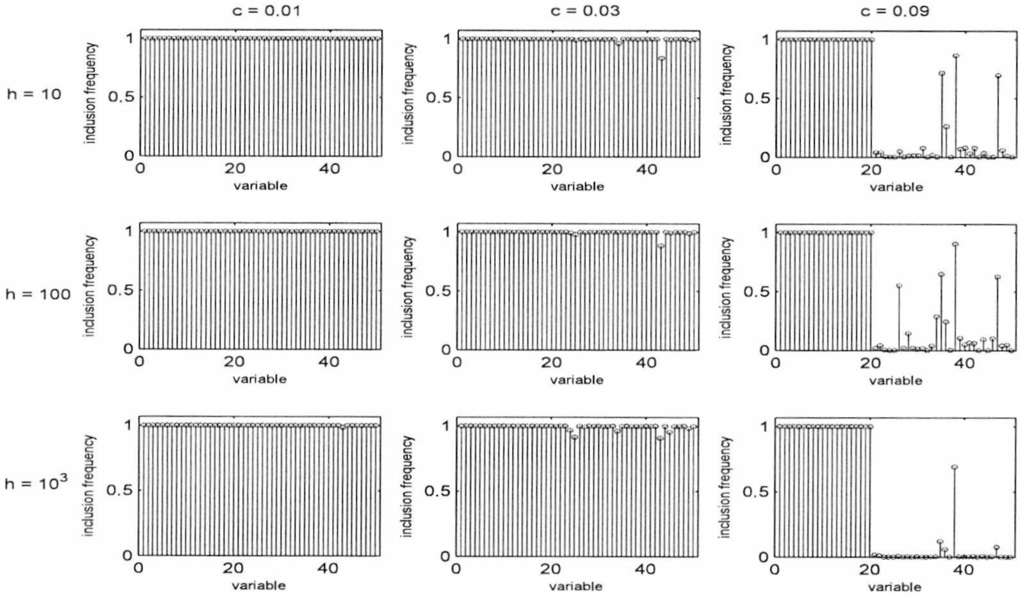

Figure 2.7.2: Simulated data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

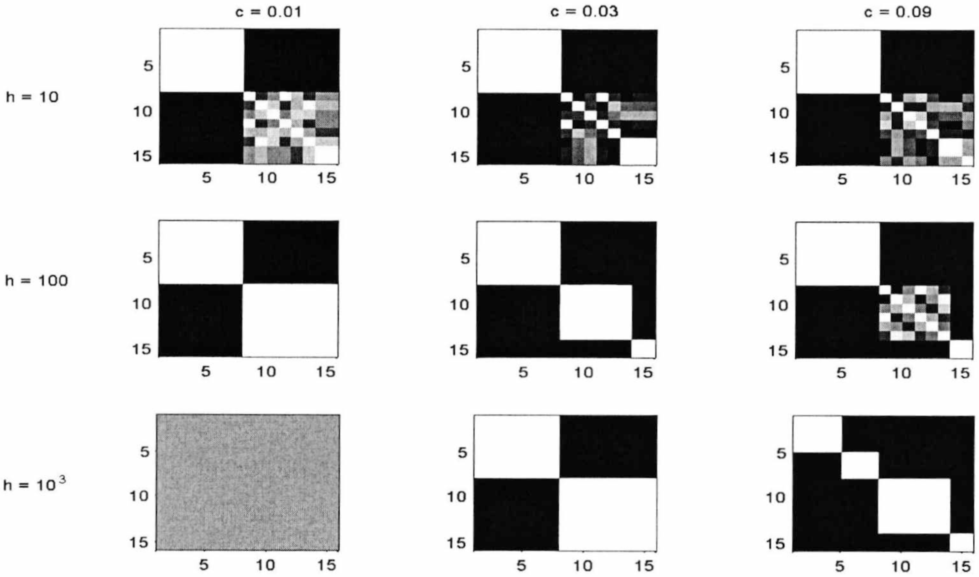

Figure 2.7.3: Simulated data: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

2.8 Conclusions

We presented a method of simultaneously selecting variables that best discriminate the sampled observations and uncovering the group structure of high-dimensional data sets with the special feature of a vast number of variables and a considerably smaller sample size. Motivated by the work of Tadesse et al. (2005), we applied a less sophisticated split/merge move under the Reversible Jump MCMC technique, that allows more flexible moves between the components, as the split or merge of empty, as well as non-empty components, is now being considered. The application of the model on a simulated data set indicated a sensitivity of the model to the choice of certain hyperparameters. Different starting points for the parameters of interest and diverse values for the various hyperparameters of the model demonstrated the need for additional care on the choice of the hyperparameters associated with the covariance structure of the model. In particular, the values of the scalars h_1, h_0, c_1 and c_0 are of great importance. The algorithm fails to identify the set of important variables and uncover the cluster structure of the data when non "suitable" values have been chosen for the crucial hyperparameters.

To overcome this deficiency and given there is no rule of thumb for the choice of these hyperparameters, we considered the idea of a two stage preprocessing procedure, stage 0 of which could facilitate the pre-specification of h_1, h_0, c_1 and c_0. Frankly, the implementation of the results showed no significant improvements. Therefore, an alternative approach on the split/merge move has been considered. Using the SAMS sampler (Dahl, 2005), we sequentially allocated the observations with probabilities conditioned on previously allocated data. The results once again were not encouraging. The algorithm still had a difficulty in properly separating the featured clusters and identifying the important variables.

Reconsidering the conclusions drawn from the implementation of the simulation study, we infer that the structure of the model and more specifically its covariance structure, needs to be altered. The correlations of the Inverse-Wishart priors on the covariance matrices of the discriminating and non-discriminating variables give the idea of evoking perplexity in already complex problems like the high-dimensional ones. Therefore, simpler structures of the covariance matrices engaged our attention as a means of overcoming such complications. A few different structures will be presented in the following chapters.

Chapter 3

Alternative Covariance Structures

3.1 Prologue

Regarding the model introduced in chapter 2, the application on a simulated data set revealed a sensitivity on the settings of the Inverse - Wishart prior distributions assigned to the covariance matrices Σ_k^D and Σ^{ND}. Along with the particular complexity characterising problems of high dimensionality, due to the vast amount of variables available opposed to the limited information provided by the small sample sizes, the correlations of the Inverse - Wishart priors impose further complications hindering the recovery of possible patterns of the data under consideration.

Motivated by the need of overcoming such adversities, we considered the investigation of different approaches for the covariance structure of the model. Starting with the simplest form of diagonal covariance matrices set proportional to the

Identity matrix, we built on this covariance structure first by replacing the Identity matrix with a diagonal matrix B. Taking the latter structure one step further and in view of controlling the increased now amount of hyperparameters, we introduce hyperpriors for the hyperparameters of the prior settings of the scalars multiplying matrix B. Being under the assumption of conjugate priors for the means μ_k^D, μ^{ND} and the covariance matrices Σ_k^D, Σ^{ND}, care on the choice of the hyperparameters h_0 and h_1 of the normal priors set on the mean vectors needs to be taken. However, we see that h_0 and h_1 can be regarded as one common factor h, the choice of which can be bypassed by giving h a prior and sampling it within the Markov chain. Having developed the structure considering a prior for h, we can now tell that we manage to get round the need of deciding values for hyperparameters such as c_1, c_0, h_1 and h_0 that we came across in the model of chapter 2. While finally, we examine the case of non-conjugacy for the priors set on the mean vectors μ_k^D, μ^{ND} and the matrices Σ_k^D, Σ^{ND}.

To conclude, we should note here that while suggesting a series of approaches for the covariance structure of the model, we examine their performance using the simulated data set initially introduced in chapter 2. Recall that the data consists of a total number of 50 variables and 15 observations. Considering 20 discriminating and 30 non-discriminating variables, the observations arise from 4 different subpopulations.

3.2 Cluster Heterogeneity

The first structure under consideration is the configuration of the covariance matrices, for both sets of discriminating and non-discriminating variables, in its simplest form. With x_i coming from the k^{th} component and being normally distributed such that,

$$x_i^D | y_i = k, \theta, \gamma \sim N\left(\mu_k^D, \Sigma_k^D\right), \quad x_i^{ND} | \theta^*, \gamma \sim N\left(\mu^{ND}, \Sigma^{ND}\right),$$

we set the covariance matrices Σ_k^D and Σ^{ND} proportional to the identity matrix. Since for the non-discriminating variables the observations are considered as a single population, for both cases of homogeneity and heterogeneity, Σ^{ND} is common along the different groups and defined as

$$\Sigma^{ND} = dI_{p-p_\gamma}.$$

A distinction between the cases of homogeneous and heterogeneous covariances for the matrices associated to the important variables needs to be made here though. In the case of homogeneity, besides the covariance matrix for the non-discriminating variables, all groups share the same covariance matrix for the set of discriminating variables as well. Hence, a single covariance matrix Σ^D proportional to the identity matrix is sufficient and therefore only a single scalar c needs to be considered. We thus write :

$$\Sigma^D = cI_{p_\gamma}.$$

On the other hand, in the case of heterogeneity $c_1, \ldots, c_G$ are being used to form the covariance matrices for each of the G subpopulations, such that,

$$\Sigma_k^D = c_k I_{p_\gamma}.$$

3.2.1 Prior Settings

Coming to the prior formulation of the model, we use settings similar to those of section 2.3. Starting with the number of discriminating variables p_γ, we consider a beta-binomial prior with expectation $pa/(a+b)$, for which the a, b parameters are chosen such that $a+b=2$, allowing a vague prior on p_γ. The number of components G follows a priori a discrete uniform on $[1, \ldots, G_{max}]$, while, a symmetric $Dirichlet\left(\alpha_w, \ldots, \alpha_w\right)$ prior is set on the vector of component weights w. Similar to the idea adopted in the model of chapter 2, in order to avoid plausible complications evoked by the changes of the dimensionality of the problem, we choose to integrate out the mean vectors of the discriminating and non-discriminating variables, μ_k^D and μ^{ND} respectively, as well as the parameters associated to the covariance matrices, (c, d) when under the assumption of homogeneity and (c_k, d) for the case of heterogeneity. To facilitate the integration, we ensured conjugacy setting Normal priors on the mean vectors,

$$\begin{aligned}
\mu_k^D | \Sigma_k^D, G &\sim N\left(\mu_0^D, h_1 \Sigma_k^D\right), \\
\mu^{ND} | \Sigma^{ND} &\sim N\left(\mu_0^{ND}, h_0 \Sigma^{ND}\right),
\end{aligned}$$

and considering the following Inverse Gamma priors for the c, c_k, d scalars :

$$c \sim IG\left(\alpha_c, \beta_c\right), \quad c_k \sim IG\left(\alpha_c, \beta_c\right), \quad d \sim IG\left(\alpha_d, \beta_d\right).$$

3.2.2 Posterior Inference

Continuing with the marginalisation over the parameters $\left(\mu_k^D, \mu^{ND}, c, d\right)$ and $\left(\mu_k^D, \mu^{ND}, c_k, d\right)$, for the cases of homogeneity and heterogeneity respectively, we can obtain the joint posterior for the parameters (y, γ, w, G). From calculations on the analytic marginalisation of the component means and covariance matrices given in Appendix B, we obtain the following joint posteriors. In the case of homogeneous covariances, we have :

$$
\begin{aligned}
f\left(X, y|G, w, \gamma\right) = {} & (2\pi)^{-np/2}\left(h_0 n+1\right)^{-(p-p_\gamma)/2}\prod_{k=1}^{G}\left[w_k^{n_k}\left(h_1 n_k+1\right)^{-p_\gamma/2}\right] \\
& \times\frac{\beta_c^{\alpha_c}}{\Gamma\left(\alpha_c\right)}\Gamma\left(\alpha_c+\frac{np_\gamma}{2}\right)\left[\beta_c+\frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{G}S_k^D\right)\right]^{-\left(\alpha_c+\frac{np_\gamma}{2}\right)} \\
& \times\frac{\beta_d^{\alpha_d}}{\Gamma\left(\alpha_d\right)}\Gamma\left(\alpha_d+\frac{n\left(p-p_\gamma\right)}{2}\right)\left[\beta_d+\frac{1}{2}\mathrm{tr}(S^{ND})\right]^{-\left(\alpha_d+\frac{n(p-p_\gamma)}{2}\right)},
\end{aligned}
\tag{3.2.1}
$$

while, under the assumption of heterogeneous covariances the joint posterior becomes :

$$
\begin{aligned}
f\left(X, y|G, w, \gamma\right) = {} & (2\pi)^{-np/2}\left(h_0 n+1\right)^{-(p-p_\gamma)/2}\prod_{k=1}^{G}\left[w_k^{n_k}\left(h_1 n_k+1\right)^{-p_\gamma/2}\right] \\
& \times\frac{\beta_d^{\alpha_d}}{\Gamma\left(\alpha_d\right)}\Gamma\left(\alpha_d+\frac{n\left(p-p_\gamma\right)}{2}\right)\left[\beta_d+\frac{1}{2}\mathrm{tr}\left(S^{ND}\right)\right]^{-\left(\alpha_d+\frac{n(p-p_\gamma)}{2}\right)} \\
& \times\frac{\beta_c^{G\alpha_c}}{\Gamma\left(\alpha_c\right)^G}\prod_{k=1}^{G}\left\{\Gamma\left(\frac{n_k p_\gamma}{2}+\alpha_c\right)\left[\beta_c+\frac{1}{2}\mathrm{tr}\left(S_k^D\right)\right]^{-\left(\alpha_c+\frac{n_k p_\gamma}{2}\right)}\right\}.
\end{aligned}
\tag{3.2.2}
$$

With $\overline{x}_k^D$ the sample mean of the variables that are included in the model for the k^{th} cluster and $\overline{x}^{ND}$ the sample mean of the non-discriminating variables, we have the matrices :

$$
S_k^D=\frac{n_k}{h_1 n_k+1}\left(\mu_0^D-\overline{x}_k^D\right)\left(\mu_0^D-\overline{x}_k^D\right)^T+\sum_{x_i^D\in C_k}\left(x_i^D-\overline{x}_k^D\right)\left(x_i^D-\overline{x}_k^D\right)^T, \tag{3.2.3}
$$

$$
S^{ND}=\frac{n}{h_0 n+1}\left(\mu_0^{ND}-\overline{x}^{ND}\right)\left(\mu_0^{ND}-\overline{x}^{ND}\right)^T+\sum_{i=1}^{n}\left(x_i^{ND}-\overline{x}^{ND}\right)\left(x_i^{ND}-\overline{x}^{ND}\right)^T. \tag{3.2.4}
$$

It is important to comment here on the computation of the above joint posteriors. For both cases of homogeneous and heterogeneous covariance matrices, we see that we are only interested in the traces of matrices S_k^D and S^{ND}. We understand therefore, that there is no necessity for generating the full matrices of $p(p+1)/2$ entries. Rather, calculations on only p elements are considered, significantly speeding up computations.

As in chapter 2, samples for the (y, γ, w, G) parameters of interest need to be drawn, and therefore MCMC techniques similar to those discussed in section 2.4 can be used. We can briefly recall. For the variable selection vector, a Metropolis search, choosing randomly between an Add, Delete or Swap move, suggests a new γ' candidate and accepts it with probability $\min\left[1, \frac{f(\gamma'|X,y,w,G)}{f(\gamma|X,y,w,G)}\right]$, where $f\left(\gamma|X, y, w, G\right)$ for the case of homogeneous covariances is given by :

$$\begin{aligned} f(\gamma|G, w, y, X) &= (h_0 n + 1)^{-(p-p_\gamma)/2} \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-p_\gamma/2} \right] \Gamma\left(\alpha_c + \frac{np_\gamma}{2}\right) \\ &\times \left[\beta_c + \frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{G} S_k^D\right)\right]^{-\left(\alpha_c + \frac{np_\gamma}{2}\right)} \Gamma\left(\alpha_d + \frac{n(p-p_\gamma)}{2}\right) \\ &\times \left[\beta_d + \frac{1}{2}\mathrm{tr}\left(S^{ND}\right)\right]^{-\left(\alpha_d + \frac{n(p-p_\gamma)}{2}\right)} \frac{B\left(p_\gamma + a, p - p_\gamma + b\right)\binom{p}{p_\gamma}}{B\left(a, b\right)}, \end{aligned} \tag{3.2.5}$$

while, for heterogeneous covariances we have,

$$\begin{aligned} f(\gamma|G, w, y, X) &= (h_0 n + 1)^{-(p-p_\gamma)/2} \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-p_\gamma/2} \right] \\ &\times \Gamma\left(\alpha_d + \frac{n(p-p_\gamma)}{2}\right) \left[\beta_d + \frac{1}{2}\mathrm{tr}\left(S^{ND}\right)\right]^{-\left(\alpha_d + \frac{n(p-p_\gamma)}{2}\right)} \\ &\times \prod_{k=1}^{G} \left\{ \Gamma\left(\alpha_c + \frac{n_k p_\gamma}{2}\right) \left[\beta_c + \frac{1}{2}\mathrm{tr}\left(S_k^D\right)\right]^{-\left(\alpha_c + \frac{n_k p_\gamma}{2}\right)} \right\} \\ &\times \frac{B\left(p_\gamma + a, p - p_\gamma + b\right)\binom{p}{p_\gamma}}{B\left(a, b\right)}. \end{aligned} \tag{3.2.6}$$

Samples for the components weights w are drawn via a Gibbs sampler from a $Dirichlet\left(\alpha_w + n_1, \ldots, \alpha_w + n_G\right)$, while the elements of the cluster allocation vector y get updated one at a time by a sub-Gibbs strategy, conditioned on the previously allocated data and using :

$$f\left(y|G, w, \gamma, X\right) = \prod_{k=1}^{G}\left[w_k^{n_k}\left(h_1 n_k + 1\right)^{-p_\gamma/2}\right]\left[\beta_c + \frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{G} S_k^D\right)\right]^{-\left(\alpha_c + \frac{n p_\gamma}{2}\right)}, \tag{3.2.7}$$

as the full conditional when under the assumption of homogeneity, and

$$\begin{aligned} f\left(y|G, w, \gamma, X\right) = & \prod_{k=1}^{G}\left[w_k^{n_k}\left(h_1 n_k + 1\right)^{-p_\gamma/2}\right]\frac{\beta_c^{G\alpha_c}}{\Gamma\left(\alpha_c\right)^G} \\ & \times \prod_{k=1}^{G}\left\{\Gamma\left(\alpha_c + \frac{n_k p_\gamma}{2}\right)\left[\beta_c + \frac{1}{2}\mathrm{tr}\left(S_k^D\right) + \right]^{-\left(\alpha_c + \frac{n_k p_\gamma}{2}\right)}\right\}, \end{aligned} \tag{3.2.8}$$

the respective to the heterogeneous case.

Finally, with the set of split/merge and birth/death moves of the RJMCMC sampler as described in section 2.4.3, we allow the search to jump between different dimensional spaces. Although the acceptance ratios given in equations (2.4.9) and (2.4.10) for the birth and death moves respectively are maintained, since the new covariance structure of the model has changed the joint posterior of the parameters (y, γ, w, G), the acceptance ratios for the split/merge moves have now been readjusted. In particular, deciding between a split or a merge move with probabilities b_G and d_G as in (2.4.2), the proposed move is being accepted with probability $\min(1, A)$. From (2.4.3), ratio A of the split move for the case of homogeneity has now become :

$$
\begin{aligned}
A = {} & \frac{\left(h_1 n_{\ell_1}+1\right)^{\frac{-p_\gamma}{2}}\left(h_1 n_{\ell_2}+1\right)^{\frac{-p_\gamma}{2}}}{\left(h_1 n_\ell+1\right)^{\frac{-p_\gamma}{2}}} \times \frac{\left\{\beta_c+\frac{1}{2}\sum_{k=1}^{G}\left[\operatorname{tr}\left(S_k^D\right)\right]\right\}^{\left(\alpha_c+\frac{np_\gamma}{2}\right)}}{\left\{\beta_c+\frac{1}{2}\sum_{k=1}^{G+1}\left[\operatorname{tr}\left(S_k^D\right)\right]\right\}^{\left(\alpha_c+\frac{np_\gamma}{2}\right)}} \\
& \times \frac{B\left(a_u,a_u\right)}{B\left(\alpha_w,G\alpha_w\right)} \times u^{\alpha_w-a_u}\left(1-u\right)^{\alpha_w-a_u} \times w_\ell^{\alpha_w} \times \frac{d_{G+1}}{b_G},
\end{aligned}
\tag{3.2.9}
$$

while, assuming heterogeneous covariances we obtain the ratio :

$$
\begin{aligned}
A = {} & \frac{\beta_c^{\alpha_c}}{\Gamma\left(\alpha_c\right)} \times \frac{\left(h_1 n_{\ell_1}+1\right)^{\frac{-p_\gamma}{2}}\left(h_1 n_{\ell_2}+1\right)^{\frac{-p_\gamma}{2}}}{\left(h_1 n_\ell+1\right)^{\frac{-p_\gamma}{2}}} \times \frac{\Gamma\left(\alpha_c+\frac{n_{\ell_1}p_\gamma}{2}\right)\Gamma\left(\alpha_c+\frac{n_{\ell_2}p_\gamma}{2}\right)}{\Gamma\left(\alpha_c+\frac{n_\ell p_\gamma}{2}\right)} \\
& \times \frac{\left[\beta_c+\frac{1}{2}\operatorname{tr}\left(S_\ell^D\right)\right]^{-\left(\alpha_c+\frac{n_\ell p_\gamma}{2}\right)}}{\left[\beta_c+\frac{1}{2}\operatorname{tr}\left(S_{\ell_1}^D\right)\right]^{-\left(\alpha_c+\frac{n_{\ell_1}p_\gamma}{2}\right)}\left[\beta_c+\frac{1}{2}\operatorname{tr}\left(S_{\ell_2}^D\right)\right]^{-\left(\alpha_c+\frac{n_{\ell_2}p_\gamma}{2}\right)}} \\
& \times \frac{B\left(a_u,a_u\right)}{B\left(\alpha_w,G\alpha_w\right)} \times u^{\alpha_w-a_u}\left(1-u\right)^{\alpha_w-a_u} \times w_\ell^{\alpha_w} \times \frac{d_{G+1}}{b_G}.
\end{aligned}
\tag{3.2.10}
$$

Reversing (3.2.9) and (3.2.10) will give us the ratios A for the acceptance probability $\min(1, A)$ of the merge move.

3.2.3 Example

We now need to examine the performance of the model with the proposed simplified covariance structure on the simulated data set suggested in section 2.5. Maintaining the vague prior on the number of discriminating variables by setting the parameters of the $Be(a,b)$ distribution such that $a+b=2$ and with the p_{prior} equal to 10, we allow G_{max} of the discrete uniform prior on the number of clusters to be equal to 15. For the Dirichlet prior on the component weights we set $\alpha_w = 1$, while the split/merge moves of the reversible jump are being proposed from a Beta with parameter $a_u = 2$. Commencing the sampler with all the variables selected as discriminating and with as many groups as the number of observations, we run 60000 iterations with a burn in of 40000. Although we have overcome the need of deciding proper values for the hyperparameters c_1 and c_0 (see section 2.5), hyperparameters h_0 and h_1 are still included in the model. Therefore, we start the analysis by exploring the effects of several different values for h_1 and h_0 in the selection of variables and the clustering task.

In Figures 3.2.1 and 3.2.2 we illustrate the number of variables included in the model and the resulting clusters for the combinations of four different values for the two hyperparameters, $(1, 10, 100, 1000)$. More specifically, in the trace plots for the variable selection task (Figure 3.2.1) we can see that although a bigger value for h_1, e.g. 100 or 1000, results in the convergence of p_γ to the expected number of 20 discriminating variables, the choice of h_0 seems to be driving the variance of the chain. This becomes more visible in the cases of $h_1 = 1$ and $h_1 = 10$. We can see that setting $h_0 = 100$ or $h_0 = 1000$ gives fairly disappointing results with the number of the selected variables varying between 20 and 40 with a big standard deviation. However, values like $h_1 = 100$ and $h_0 = 1000$ can overcome the effect of h_0 and derive trace plots that show convergence to the desired number of important variables. Generally, if we look across the columns of Figure 3.2.1 we can see that no matter the choice of h_0, a good value for h_1 (e.g $100, 1000$) can lead to convergence of p_γ with a small standard deviation. On the other hand, looking along the rows we can see that the smaller the value of h_1 the stronger the impact of h_0 in the analysis ($h_1 = 1$, $h_1 = 10$).

The importance of h_1 becomes particularly clear observing Figure 3.2.2. Looking across the columns, we can see how h_1 drives the algorithm into forming the four groups. Care needs to be given on the choice of h_1 though; while with $h_1 = 1$ there is a difficulty in splitting the first two groups, the four groups can be successfully extracted for $h_1 = 10$, but again under the case of $h_0 = 1000$, the algorithm

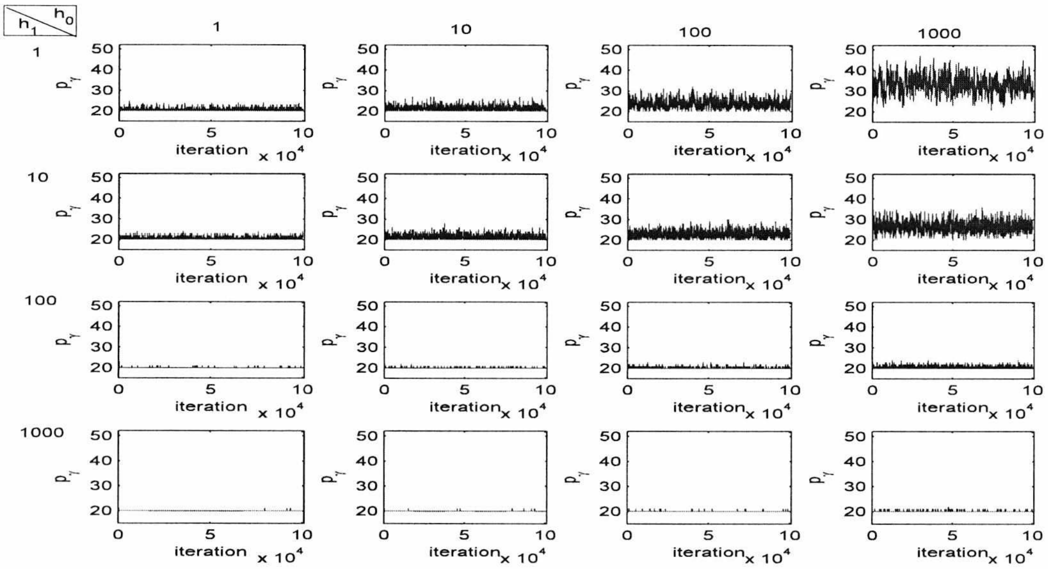

Figure 3.2.1: Cluster Heterogeneity for the simulated data set: Trace plots of the total number of discriminating variables included in the model, p_γ, for different combinations of the hyperparameters h_1, h_0, letting $\alpha_c = \beta_c = \alpha_d = \beta_c = 2$ and assuming unequal covariances.

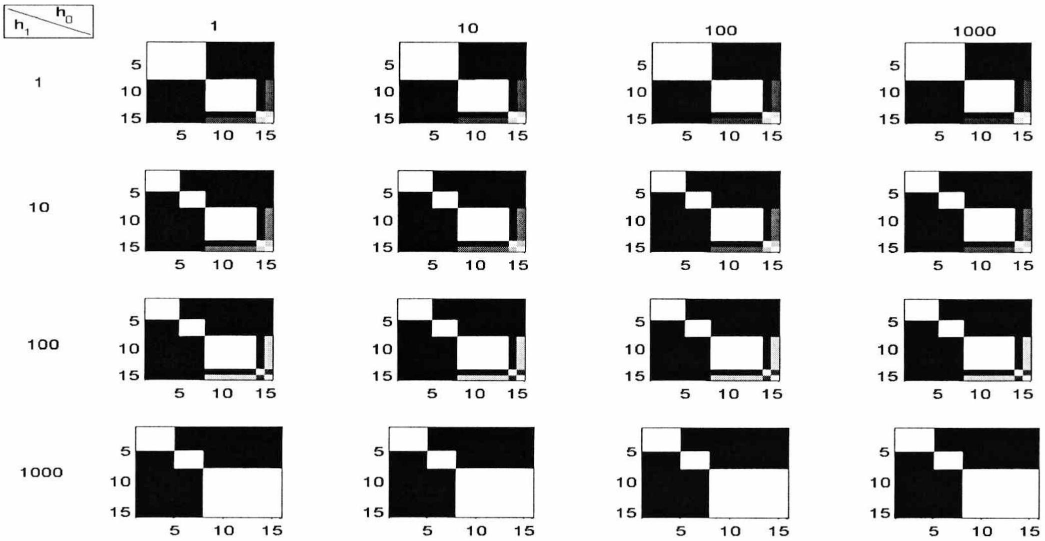

Figure 3.2.2: Cluster Heterogeneity for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for different combinations of the hyperparameters h_1, h_0, letting $\alpha_c = \beta_c = \alpha_d = \beta_c = 2$ and assuming unequal covariances.

recovers only three groups. Meanwhile, we observe that the choice of h_0 does not have a strong impact in the formulation of the groups.

Concluding the small investigation on the choice of h_1 and h_0, we can summarise our remarks in the following. It looks, from the resulting trace plots of p_γ and the clustering maps, as though the choice of h_1 is more important than that of h_0. Since the value of h_1 is what we focus on, the idea of setting h_0 equal to h_1, as we did in our examples in chapter 2, seems utterly reasonable. Therefore, we can assume a single hyperparameter, say h $(h = h_1 = h_0)$.

Now, setting $h = 10, 100, 1000$, we examine the cases of different values for the hyperparameters of the $IG(a_c, b_c)$ and $IG\left(a_d, b_d\right)$ priors on the scalars c_k's and d. As indicated in Figures 3.2.3 - 3.2.6, we tried the values $1, 2$ and 1000 for the hyperparameters $\alpha_c, \beta_c, \alpha_d, \beta_d$ and checked the results across the different values of h. The trace plots of p_γ, for all cases, show a convergence to the 20 discriminating variables, with the cases of higher values of h indicating less variation. Although the histogram of the number of selected variables (Figure 3.2.4), for the case $\alpha_c = \beta_c = \alpha_d = \beta_d = 1000$ and $h = 10$, indicates the selection of roughly 23 variables, Figure 3.2.5 shows that the main core of the 20 variables are always being selected, with $h = 10$ assigning higher probability to an additional small number of variables.

Finally, looking at the clustering maps (Figure 3.2.6), we can tell that the choice of $\alpha_c, \beta_c, \alpha_d$ and β_d can affect the algorithm, with $\alpha_c = \beta_c = \alpha_d = \beta_d = 1$ fully recovering the four clusters regardless the choice of h, while setting $\alpha_c = \beta_c = \alpha_d = \beta_d = 2$ and increasing the value of h can gradually lead to the formulation of three groups and in particular the merge of the third and fourth cluster.

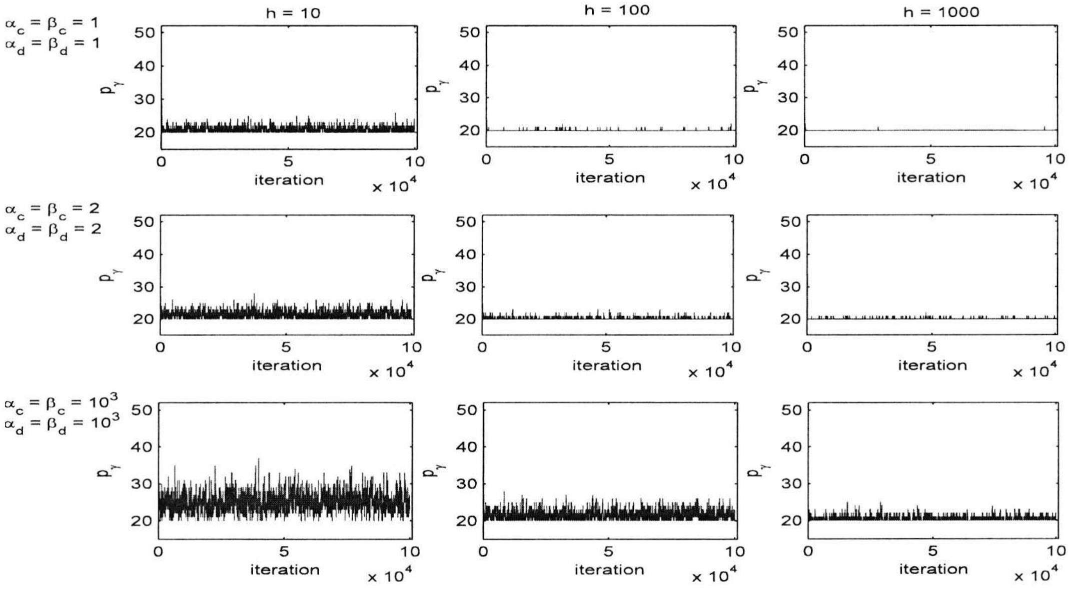

Figure 3.2.3: Cluster Heterogeneity for the simulated data set: Trace plots of the total number of discriminating variables included in the model, p_γ, for $h = 10, 100, 1000$ and and $\alpha_c = \beta_c = \alpha_d = \beta_d$ equal to $1, 2,$ and 1000, under the assumption of unequal covariance matrices.

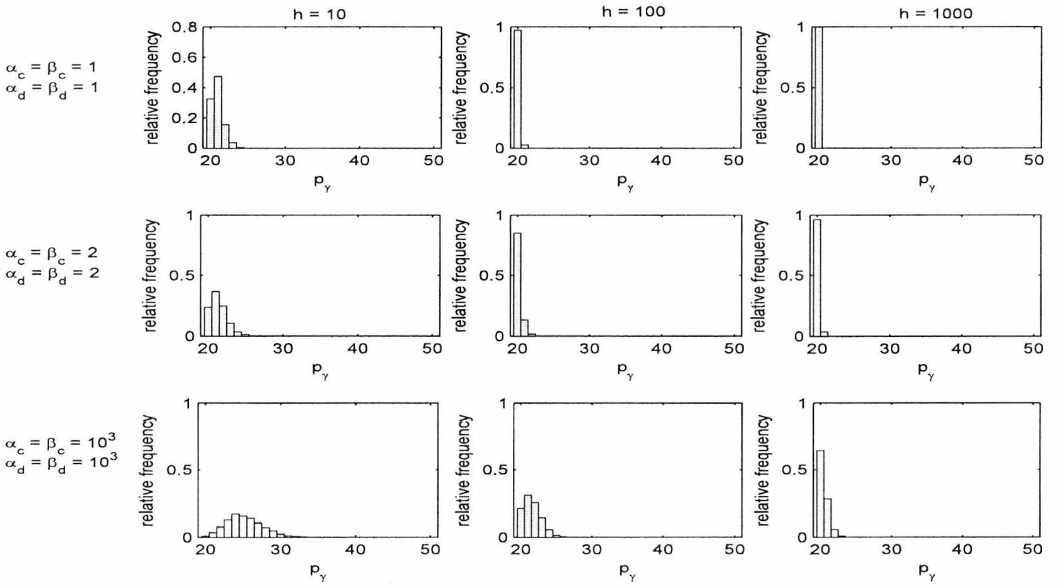

Figure 3.2.4: Cluster Heterogeneity for the simulated data set: Histograms of the total number of discriminating variables, p_γ, assuming unequal covariance matrices, for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ with values $1, 2,$ and 1000.

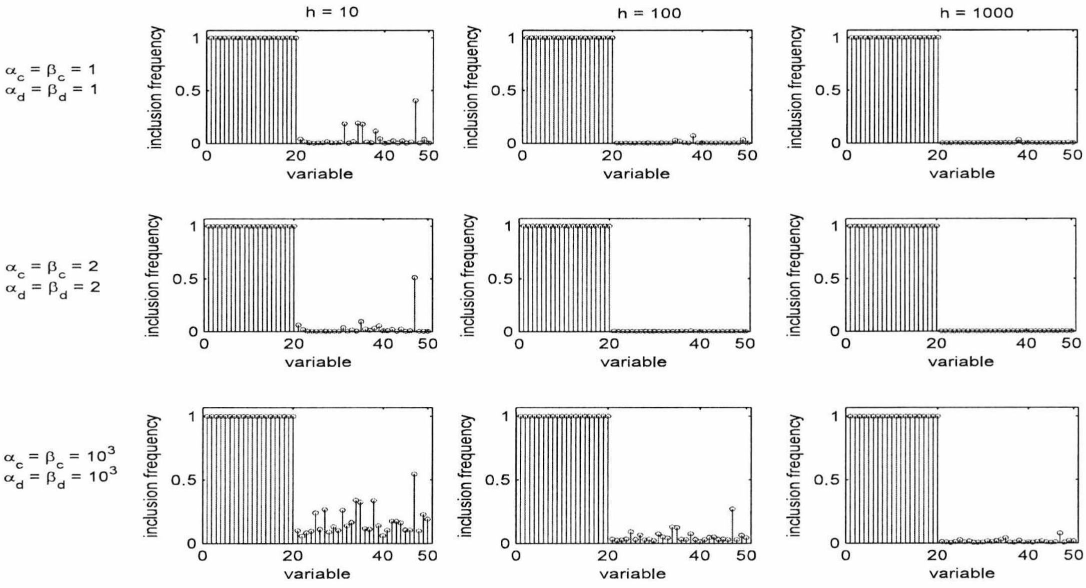

Figure 3.2.5: Cluster Heterogeneity for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, assuming unequal covariance matrices, for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ with values $1, 2,$ and 1000.

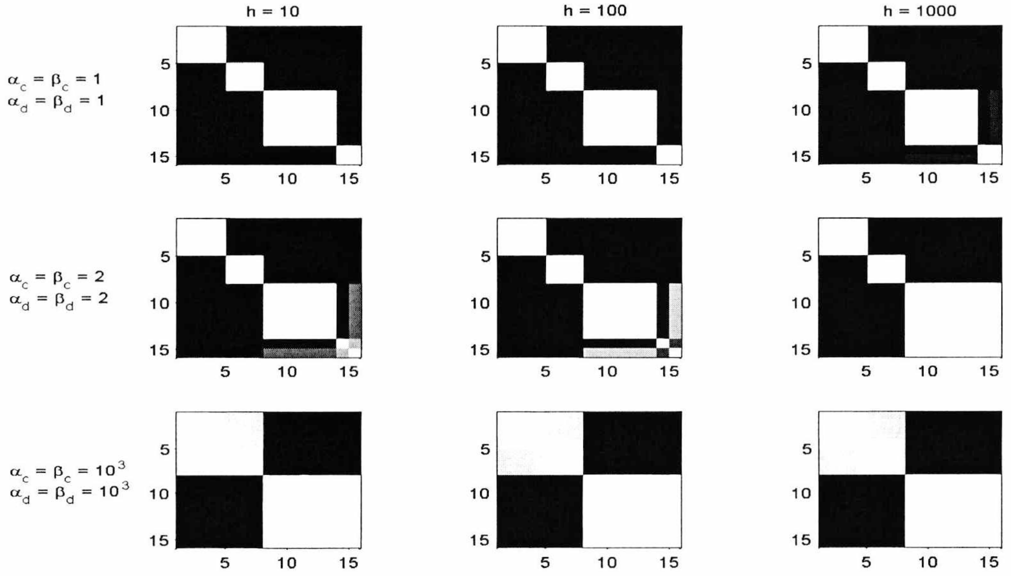

Figure 3.2.6: Cluster Heterogeneity for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ equal to $1, 2,$ and 1000, under the assumption of unequal covariance matrices.

3.3 Cluster-Variable Heterogeneity

Taking the covariance structure of the Cluster Heterogeneity (CH) model one step further, we now regard the matrices Σ_k^D and Σ^{ND} proportional to a diagonal matrix. With the normally distributed observations :

$$x_i^D | y_i = k, \theta, \gamma \sim N\left(\mu_k^D, \Sigma_k^D\right), \quad x_i^{ND} | \theta^*, \gamma \sim N\left(\mu^{ND}, \Sigma^{ND}\right),$$

and with the variances of the p covariates, σ_j^2, $j = 1, ..., p$ being the elements of the diagonal matrix B_p, i.e.

$$B = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \sigma_p^2 \end{pmatrix},$$

we define the covariance matrices of the discriminating variables,

$$\Sigma^D = cB_{p_\gamma}, \quad \text{and} \quad \Sigma_k^D = c_k B_{p_\gamma},$$

for assuming homogeneity and heterogeneity along the groups respectively; while, for the non-discriminating variables, we have the covariance matrix,

$$\Sigma^{ND} = B_{p-p_\gamma},$$

which maintains its form under both cases. Evidently, we note that B_{p_γ} and B_{p-p_γ} are submatrices of B for the variances corresponding to the discriminating and non-discriminating variables respectively.

3.3.1 Prior Settings

The prior settings for the Cluster-Variable Heterogeneity model, resemble the settings we have discussed so far with the additional prior formulation of the variances σ_j^2. Along with the known by now beta-binomial prior on the number of discriminating variables p_γ, as well as the discrete uniform prior on $[1, \ldots, G_{max}]$ for the number of components G, we retain the symmetric *Dirichlet* $(\alpha_w, \ldots, \alpha_w)$ prior on the component weights and hire conjugate priors for the mean vectors and the scalars of the covariance matrices. More specifically, we have :

$$\mu_k^D|\Sigma_k^D, G \sim N\left(\mu_0^D, h_1\Sigma_k^D\right),$$
$$\mu^D|\Sigma^{ND} \sim N\left(\mu_0^{ND}, h_0\Sigma^{ND}\right),$$

where, for the scalars of the Σ_k^D and Σ^{ND} covariance matrices, we use Gamma priors such that :

$$c \sim Ga\left(\alpha_c, \beta_c\right), \quad c_k \sim Ga\left(\alpha_c, \beta_c\right), \quad d \sim Ga\left(\alpha_d, \beta_d\right).$$

Finally, for the p diagonal elements of B (or of B_{p_γ} and B_{p-p_γ}), we a priori draw the precisions σ_j^{-2}'s independently from Gamma distributions with parameters α_s, β_s, i.e. $\sigma_j^{-2} \sim Ga\left(\alpha_s, \beta_s\right)$.

3.3.2 Posterior Inference

Examining the posterior formulation of the model, we certainly cannot overlook the change of dimensionality imposed by the variable selection task as well as the sampling of the number of clusters and the cluster allocations. The conjugacy of the priors on the mean vectors allows us to integrate out the parameters μ_k^D and μ^{ND}, however, as the algorithm jumps between dimensions, although under the assumption of homogeneous covariances σ_j^2's get affected from the change of dimensionality, in the case of heterogeneity the dimension of both sets of parameters, σ_j^2's and c_k's, needs to be updated. Therefore, the parameters of interest and the sampling techniques used, differ for the two cases of homogeneous an heterogeneous covariances.

Starting with the assumption of equal variances across the G groups, being under the Reversible Jump chain, it is difficult to propose the precisions σ_j^{-2}, while an additional time burden would be imposed in the algorithm. Therefore, since marginalisation of the joint posterior over σ_j^{-2}'s can be achieved, we choose to integrate the precisions out. Along with the integration over the mean vectors μ_k^D and μ^{ND}, the calculations result to a joint posterior of the parameters (y, γ, w, G, c) with the following form :

$$\begin{aligned} f\left(X, y | G, w, \gamma, c\right) = & \left(2\pi\right)^{-np/2} \frac{\beta_s^{\alpha_s p}}{\Gamma\left(\alpha_s\right)^p} \Gamma\left(\alpha_s + \frac{n}{2}\right)^p \left(h_0 n + 1\right)^{-(p-p_\gamma)/2} c^{-np_\gamma/2} \\ & \times \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-p_\gamma/2} \right] \prod_{ND} \left\{ \beta_s + \frac{1}{2} \left[S^{ND}\right]_{jj} \right\}^{-\left(\alpha_s + \frac{n}{2}\right)} \\ & \times \prod_{D} \left\{ \beta_s + \frac{1}{2}\frac{1}{c} \left[\sum_{k=1}^{G} S_k^D \right]_{jj} \right\}^{-\left(\alpha_s + \frac{n}{2}\right)}, \end{aligned} \tag{3.3.1}$$

under which, calculations on only the p variances of the S_k^D, S^{ND} matrices is needed.

Looking at the sampling techniques for the parameters of interest for the homogeneous covariances case, (γ, w, y, G, c), we start with the sampling of the γ vector using a Metropolis search. With one of the Add/Delete/Swap moves suggesting a new candidate vector and with full conditional :

$$f\left(\gamma | G, w, y, X, c\right) \propto f\left(X, y | G, w, \gamma, c\right) p\left(\gamma | G\right), \tag{3.3.2}$$

we accept the proposed γ' with probability min $\left[1, \frac{f(\gamma'|X,y,w,G)}{f(\gamma|X,y,w,G)}\right]$.

In the next step, samples of the component weights w are being drawn from a Dirichlet distribution with parameters $(\alpha_w + n_1, \ldots, \alpha_w + n_G)$ using a Gibbs sampler. While, on the following, we update the elements of the allocation vector y one at a time via a sub - Gibbs strategy with probability conditioned on the previous allocations,

$$f\left(y_i = k|X, y_{(-i)}, \gamma, w, G\right) \propto f\left(X, y_i = k, y_{(-i)}|G, w, \gamma\right), \tag{3.3.3}$$

where, $f\,(y|G, w, \gamma, X, c) \propto f\,(X, y|G, w, \gamma, c)\,.$

Before performing the RJMCMC sampler, we introduce the sampling of the scalar c. Suggesting a new c from its prior distribution, we update c using the Metropolis-Hastings algorithm, accepting the new candidate with probability :

$$\alpha = \min\left\{1, \frac{\pi\,(c'|G, w, \gamma, X, y)}{\pi\,(c|G, w, \gamma, X, y)}\right\},$$

where, $\pi\,(c'|G, w, \gamma, X, y)$ is the full conditional of c as in :

$$f\,(c|G, w, \gamma, X, y) \propto f\,(X, y|G, w, \gamma, c)\,p\,(c)\,. \tag{3.3.4}$$

Finally, the algorithm concludes with the split/merge and birth/death moves of the Reversible Jump sampler. Using the b_G and d_G probabilities as defined in (2.4.2), a split or a merge move is being proposed. Assuming a split of an empty or non-empty component has been suggested, the move is being accepted with probability $\min(1, A)$, for which ratio A has now been reformed into :

$$A = \frac{(h_1 n_{\ell_1} + 1)^{-p_\gamma/2} (h_1 n_{\ell_2} + 1)^{-p_\gamma/2}}{(h_1 n_\ell + 1)^{-p_\gamma/2}} \times \frac{\prod_{D} \left\{ \beta_s + \frac{1}{2}\frac{1}{c} \left[\sum_{k=1}^{G+1} S_k^D \right]_{jj} \right\}^{-\left(\alpha_s + \frac{n}{2}\right)}}{\prod_{ND} \left\{ \beta_s + \frac{1}{2}\frac{1}{c} \left[\sum_{k=1}^{G} S_k^D \right]_{jj} \right\}^{-\left(\alpha_s + \frac{n}{2}\right)}}$$
$$\times \frac{B(a_u, a_u)}{B(\alpha_w, G\alpha_w)} \times u^{\alpha_w - a_u} (1-u)^{\alpha_w - a_u} \times w_\ell^{\alpha_w} \times \frac{d_{G+1}}{b_G}. \qquad (3.3.5)$$

We understand that the reverse of (3.3.5) will give us the corresponding ratio for the merging of two components, while the sampler will conclude with the generation or deletion of an empty component. Proceeding into a birth or a death move with b_{G_0} and d_{G_0} probabilities as in (2.4.7), a suggested birth move will be accepted with probability (2.4.9), while (2.4.10) will accept a death move.

Having examined the parameters of interest along with their joint posterior and the sampling techniques for the case of homogeneous covariances, we now need to look at the assumption of unequal covariances across the groups. This time, as mentioned earlier, both sets of σ_j^{-2}'s and the vector of c_k's changes dimension as the sampler jumps between dimensional spaces; however, the integration of both sets is not feasible. Reconsidering the RJMCMC sampler, as we split/merge and generate/delete new components, the dimension of the vector with the scalars $c_1, \ldots, c_G$, needs to comply with the proposed moves, accompanied with proper proposals constructing the scalars assigned to the new components. We understand that such a choice would impede the efficiency of the sampler and we therefore decide to integrate out scalars c_k's and in return sample the precisions σ_j^{-2}.

After integrating out the mean vectors μ_k^D, μ^{ND} and the c_k's, from a joint posterior of the form :

$$f\left(X, y|G, w, \gamma, \sigma_j^{-2}\right) = \int f\left(X, y|\gamma, G, w, \mu^D, \mu^{ND}, \sigma_j^{-2}, c_k\right) \times f\left(c_k|G\right) f\left(\mu^D|G, \gamma, \Sigma^D\right)$$
$$\times f\left(\Sigma^D|G, \gamma\right) f\left(\mu^{ND}|\Sigma^{ND}, \gamma\right) f\left(\Sigma^{ND}|\gamma\right) d\mu^D\, d\mu^{ND}\, dc_k,$$

with the parameters of interest being $\left(\gamma, w, y, G, \sigma_j^{-2}\right)$, we obtain the following joint posterior :

$$\begin{aligned} f\left(X, y|G, w, \gamma, \sigma_j^{-2}\right) = {} & (2\pi)^{-np/2} \left(h_0 n + 1\right)^{-(p-p_\gamma)/2} \frac{\beta_c^{G\alpha_c}}{\Gamma\left(\alpha_c\right)^G} \left|B_p^{-1}\right|^{n/2} \\ & \times \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-p_\gamma/2}\right] \exp\left[-\frac{1}{2} tr\left(B_{p-p_\gamma}^{-1} S^{ND}\right)\right] \\ & \times \prod_{k=1}^{G} \left\{ \frac{2K_{\left(\alpha_c - \frac{n_k p_\gamma}{2}\right)}\left(\sqrt{2\beta_c tr\left(B_{p_\gamma}^{-1} S_k^D\right)}\right)}{\left[\frac{2\beta_c}{tr\left(B_{p_\gamma}^{-1} S_k^D\right)}\right]^{\frac{\alpha_c - \frac{n_k p_\gamma}{2}}{2}}} \right\}, \end{aligned} \tag{3.3.6}$$

where, $K_{\left(\alpha_c - \frac{n_k p_\gamma}{2}\right)}\left(\sqrt{2\beta_c tr\left(B_{p_\gamma}^{-1} S_k^D\right)}\right)$ is the modified Bessel function of the second kind and S_k^D, S^{ND} as defined in (3.2.3) and (3.2.4) respectively. Once again, traces indicate the need for only p calculations.

The algorithm starts with the sampling of the γ vector with the usual Add/Delete /Swap moves of the Metropolis search, accepting the proposed vector γ with probability min $\left[1, \frac{f(\gamma'|X,y,w,G)}{f(\gamma|X,y,w,G)}\right]$, the full conditional of which is given by :

$$f\left(\gamma|G, w, y, X, \sigma_j^{-2}\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}\right) p\left(\gamma|G\right). \tag{3.3.7}$$

Coming to the cluster allocations and the component weights, a Gibbs sampler, with a *Dirichlet* $\left(\alpha_w + n_1, \ldots, \alpha_w + n_G\right)$ posterior distribution, delivers the samples of w, while, similarly to the homogeneous case, the new cluster membership of the observations indicated by y, are being updated one element at a time via a sub - Gibbs strategy with probability conditioned on the previous allocations, recall,

$$f\left(y_i = k|X, y_{(-i)}, \gamma, w, G\right) \propto f\left(X, y_i = k, y_{(-i)}|G, w, \gamma\right), \tag{3.3.8}$$

where this time the full conditional for vector y is given by

$$f\left(y|G, w, \gamma, X, \sigma_j^{-2}\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}\right), .$$

We proceed with the sampling of the precisions σ_j^{-2}'s. With the full conditional of σ_j^{-2}'s being of no closed form, i.e.

$$f\left(\sigma_j^{-2}|G, w, y, X, \gamma\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}\right) p\left(\sigma_j^{-2}\right), \tag{3.3.9}$$

a standard approach for drawing samples of σ_j^{-2} would be a Metropolis - Hastings algorithm. However, we choose to draw σ_j^{-2}'s via a variant of the Gibbs sampler, the Collapsed-Gibbs sampler. Although samples of the parameters γ, y and w are being drawn having marginalised the posterior over the scalars c_k's, the idea of Collapsed-Gibbs sampler introduces the sampling of σ_j^{-2}'s having blocked on c_k's. More specifically, while c_k's are no longer being considered in the analysis, i.e. we do not consider sampling them since we chose to integrate them out for reasons we explained earlier in this section, the samples of the variances σ_j^{-2}'s are being drawn from their full conditional being conditioned on c_k's, i.e. $f\left(\sigma_j^{-2}|G, w, y, X, \gamma, c_k\right)$. In our case, that would be :

$$f\left(\sigma_j^{-2}|G, w, y, X, \gamma, c_k\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c_k\right) p\left(\sigma_j^{-2}\right), \tag{3.3.10}$$

where, $f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c_k\right)$ is the joint posterior of $\left(G, w, y, \gamma, \sigma_j^{-2}, c_k\right)$ and $p\left(\sigma_j^{-2}\right)$ the Gamma prior of the p variances. However, (3.3.10) yields a Gamma posterior such that :

$$\sigma_j^{-2} \sim Ga\left(\alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2}\left[\sum_{k=1}^{G} \frac{1}{c_k} S_k^D\right]_{jj}\right), \tag{3.3.11}$$

for the variances corresponding to the p_γ discriminating variables, and a Gamma posterior of the form :

$$\sigma_j^{-2} \sim Ga\left(\alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2} S_{jj}^{ND}\right), \tag{3.3.12}$$

for the $(p - p_\gamma)$ variances of the non - discriminating variables. We understand that the posterior of the σ_j^{-2}'s is of closed form and therefore a Gibbs sampler can be applied. Nonetheless, before performing the sampling of σ_j^{-2}'s, we need to draw samples for the scalars $c_k, k = 1, \ldots, G$. With a full conditional of the general form :

$$f\left(c_k | G, w, y, X, \gamma, \sigma_j^{-2}\right) \propto f\left(X, y | G, w, \gamma, \sigma_j^{-2}, c_k\right) p\left(c_k\right), \tag{3.3.13}$$

each of the G c_k's is being sampled, via again a Gibbs sampler, from a Generalised Inverse Gaussian, such that :

$$c_k \sim GIG\left(p_c, a_c, b_c\right), \tag{3.3.14}$$

with $p_c = \alpha_c - n_k p_\gamma / 2$, $a_c = 2\beta_c$ and $b_c = tr\left(B_{p_\gamma}^{-1} S_k^D\right)$. Having completed the Collapsed-Gibbs step, the algorithm continues with the Reversible Jump MCMC technique, considering once again the posterior marginalised over the c_k's, i.e. $f\left(X, y | G, w, \gamma, \sigma_j^{-2}\right)$. Similarly to the homogeneous case, we have a split or a merge move proposed and accepted with probability $\min(1, A)$. For the split move ratio A takes the form :

$$\begin{aligned} A = & \frac{\beta_c^{\alpha_c}}{\Gamma\left(\alpha_c\right)} \times \frac{\left(h_1 n_{\ell_1}+1\right)^{-p_\gamma/2}\left(h_1 n_{\ell_2}+1\right)^{-p_\gamma/2}}{\left(h_1 n_\ell+1\right)^{-p_\gamma/2}} \\ & \times \frac{2K_{\left(\alpha_c-\frac{n_{\ell_1}p_\gamma}{2}\right)}\left(\sqrt{2\beta_c tr\left(B_{p_\gamma}^{-1}S_{\ell_1}^D\right)}\right) 2K_{\left(\alpha_c-\frac{n_{\ell_2}p_\gamma}{2}\right)}\left(\sqrt{2\beta_c tr\left(B_{p_\gamma}^{-1}S_{\ell_2}^D\right)}\right)}{2K_{\left(\alpha_c-\frac{n_\ell p_\gamma}{2}\right)}\left(\sqrt{2\beta_c tr\left(B_{p_\gamma}^{-1}S_\ell^D\right)}\right)} \\ & \times \frac{\left(\frac{2\beta_c}{tr\left(B_{p_\gamma}^{-1}S_\ell^D\right)}\right)^{\left(\alpha_c-\frac{n_\ell p_\gamma}{2}\right)/2}}{\left(\frac{2\beta_c}{tr\left(B_{p_\gamma}^{-1}S_{\ell_1}^D\right)}\right)^{\left(\alpha_c-\frac{n_{\ell_1}p_\gamma}{2}\right)/2}\left(\frac{2\beta_c}{tr\left(B_{p_\gamma}^{-1}S_{\ell_2}^D\right)}\right)^{\left(\alpha_c-\frac{n_{\ell_2}p_\gamma}{2}\right)/2}} \\ & \times \frac{B\left(a_u,a_u\right)}{B\left(\alpha_w,G\alpha_w\right)} \times u^{\alpha_w-a_u}\left(1-u\right)^{\alpha_w-a_u} \times w_\ell^{\alpha_w} \times \frac{d_{G+1}}{b_G}, \end{aligned} \qquad (3.3.15)$$

while, for the merge move, ratio A will correspond to the reverse of (3.3.15). Finally, we have the birth and the death moves, the acceptance probability of which is $\min\left(1, A\right)$, with A as defined in (2.4.9) and (2.4.10) respectively.

3.3.3 Example

Finishing the construction of the model with the new covariance structure, its application on our usual simulated data set of the 50 variables and the 15 observations assuming heterogeneity is essential. Once again we iterated our runs 60000 times using a burn in of 40000 iterations and set $p_{prior} = 10$, $a+b=2$, $G_{max}=15$, $\alpha_w=1$, and $a_u=2$. For the so far crucial hyperparameters h_1 and h_0, we considered the values 100 and 1000, while this time a few new hyperparameters have been introduced in the algorithm, that is α_c, β_c and α_s, β_s. Care about the impact of the new hyperparameters on the convergence of the selected variables and the formed clusters has been taken by checking the resulting histograms and cluster matrices for certain different values.

As far as α_s and β_s of the Gamma prior on the precisions σ_j^{-2}'s are concerned, we tried the pairs $(2,1)$, $(3,2)$ and $(5,4)$, allowing precisions with expected value in the interval $[1,2]$. However, regarding the hyperparameters of the Gamma prior used on c_k's, i.e. α_c, β_c, we thought as follows. To begin with, we are interested in examining the within covariance structure of the groups with respect to the overall, and more specifically, to be able to identify the cluster structure of the observations, we understand that the within groups variance needs to be smaller than the overall. Therefore we expect $c_k \in (0,1)$. Consequently, values of (α_c, β_c) such that the expected a priori value of a scalar c_k is $\frac{1}{2}$ or $\frac{1}{4}$ would be a reasonable choice. We experimented with the pairs $(1,4)$, $(3,12)$, $(1,2)$, etc (see Figures for the full list of (α_c, β_c) pairs) for the three pairs of (α_s, β_s) $[(2,1)$, $(3,2)$ and $(5,4)]$ and allowing $h_1 = h_0 = 100$ and $h_1 = h_0 = 1000$.

Looking at Figures 3.3.1 - 3.3.6, one can interpret the importance of the choice of h_1 and h_0. Having $h_0 = h_1 = 100$, results to the selection of the 20 discriminating variables originally defined as the discriminating ones (Figure 3.3.2) and the recovery of the four clusters of the observations, for all combinations of the (α_s, β_s) and (α_c, β_c) pairs, except the case of $(\alpha_s, \beta_s) = (2,1)$ with $(\alpha_c, \beta_c) = (1,4)$. Setting $h_0 = h_1 = 1000$, on the other hand, with the only exception of $(\alpha_c, \beta_c) = (4,8)$ for which both the discriminating variables and the cluster structure of the data have been successfully recovered, we observe that the pair (α_s, β_s) seems to be of importance. With $(\alpha_s, \beta_s) = (2,1)$ resulting to a model favouring the entire set of variables as discriminating, failing to recover the true cluster structure of the data, we note that a change on the values on the α_s and β_s hyperparameters can improve the results dragging the algorithm into convergence to the correct set of the 20 discriminating variables and the structure of four groups of observations [pairs $(3,2)$ and $(5,4)$].

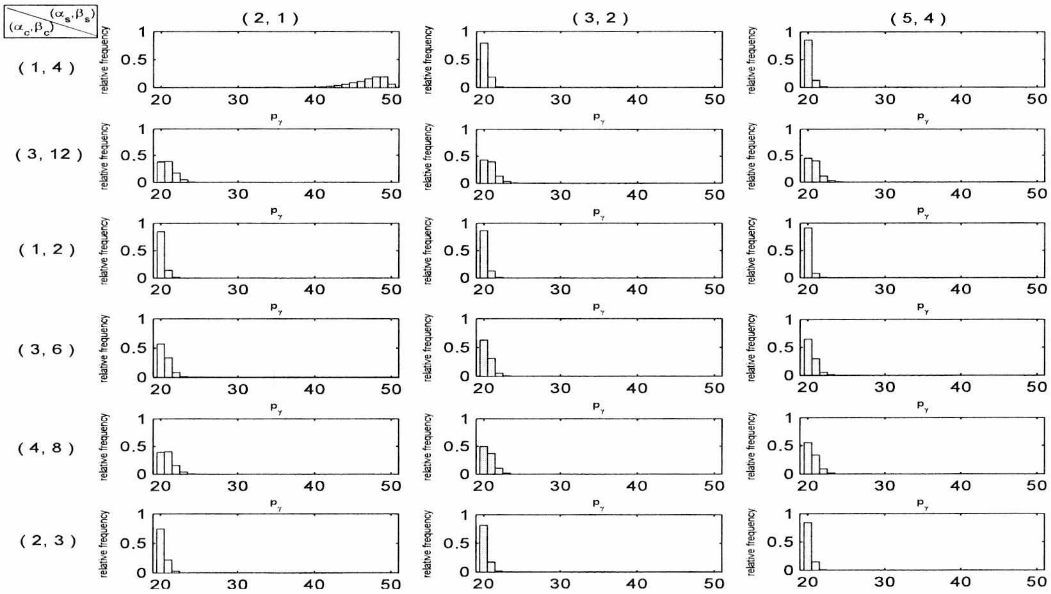

Figure 3.3.1: Cluster-Variable Heterogeneity for the simulated data set: Histograms of the total number of discriminating variables, p_γ, for various pairs of (α_c, β_c), (α_s, β_s), assuming heterogeneity and setting $h = 100$.

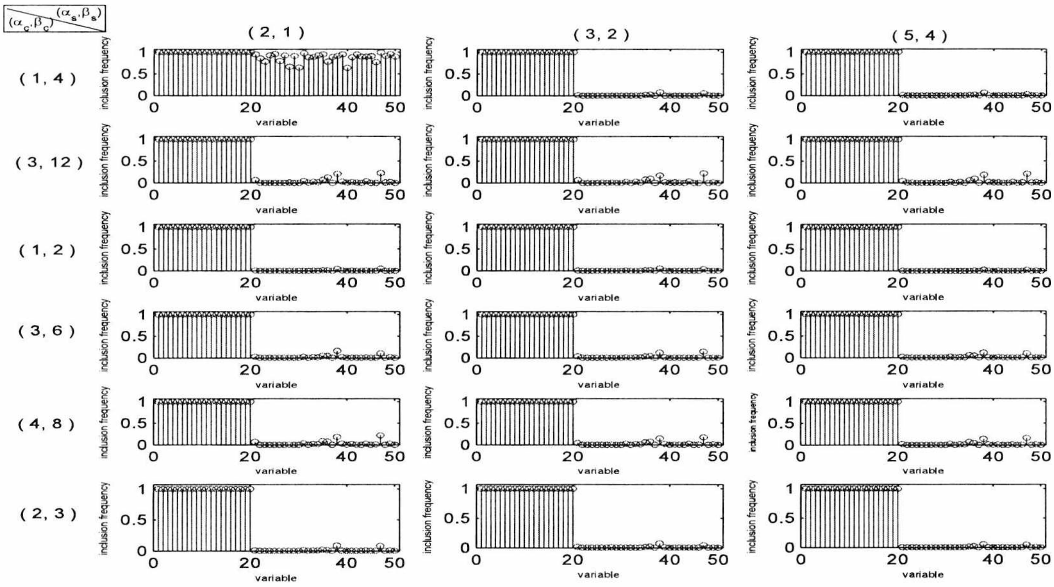

Figure 3.3.2: Cluster-Variable Heterogeneity for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various pairs of (α_c, β_c), (α_s, β_s), assuming heterogeneity and setting $h = 100$.

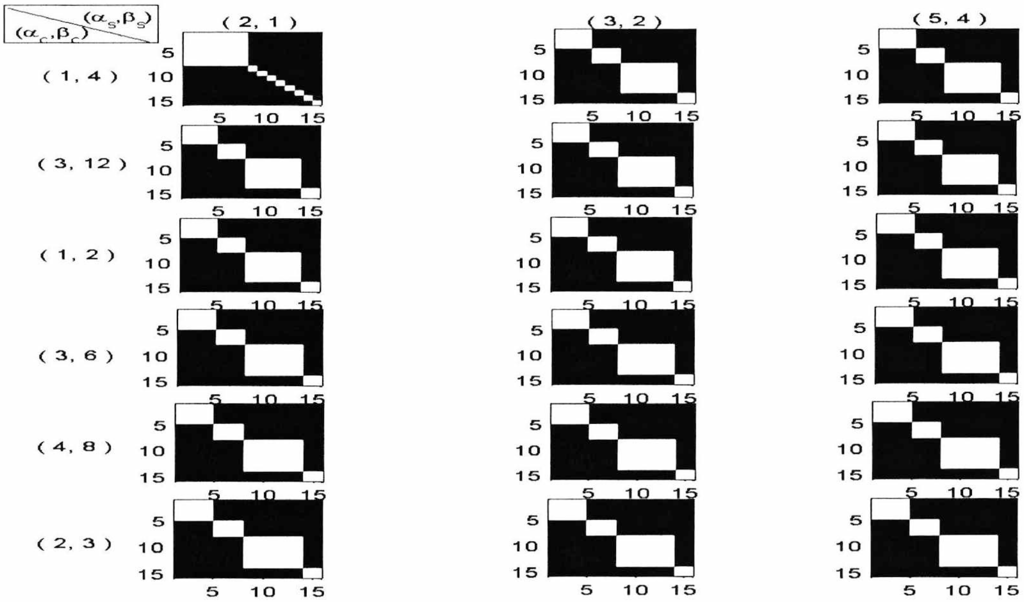

Figure 3.3.3: Cluster-Variable Heterogeneity for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various pairs of (α_c, β_c), (α_s, β_s), assuming heterogeneity and setting $h = 100$.

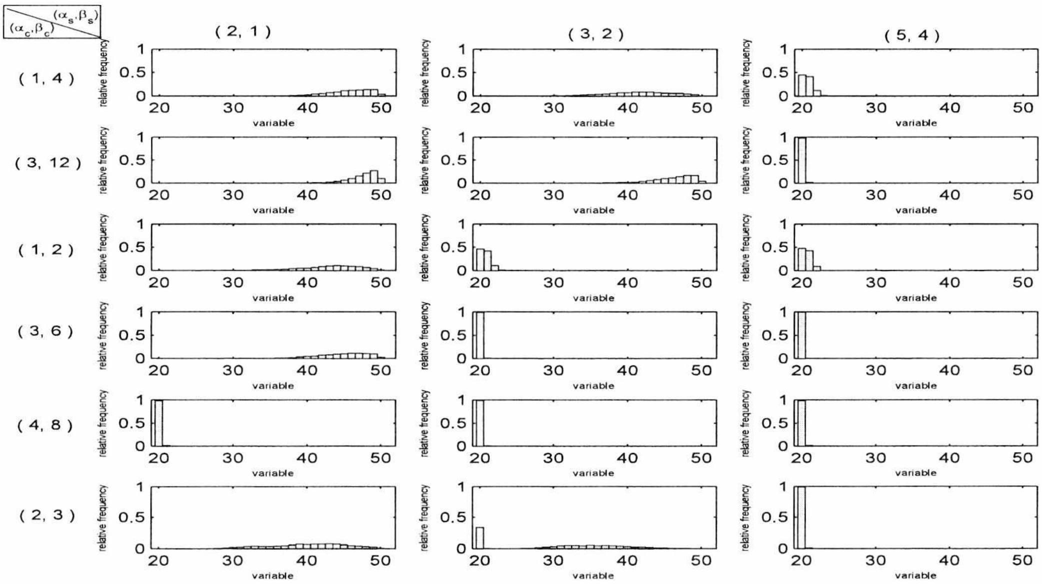

Figure 3.3.4: Cluster-Variable Heterogeneity for the simulated data set: Histograms of the total number of discriminating variables,. p_γ, for various pairs of (α_c, β_c), (α_s, β_s), assuming heterogeneity and setting $h = 1000$.

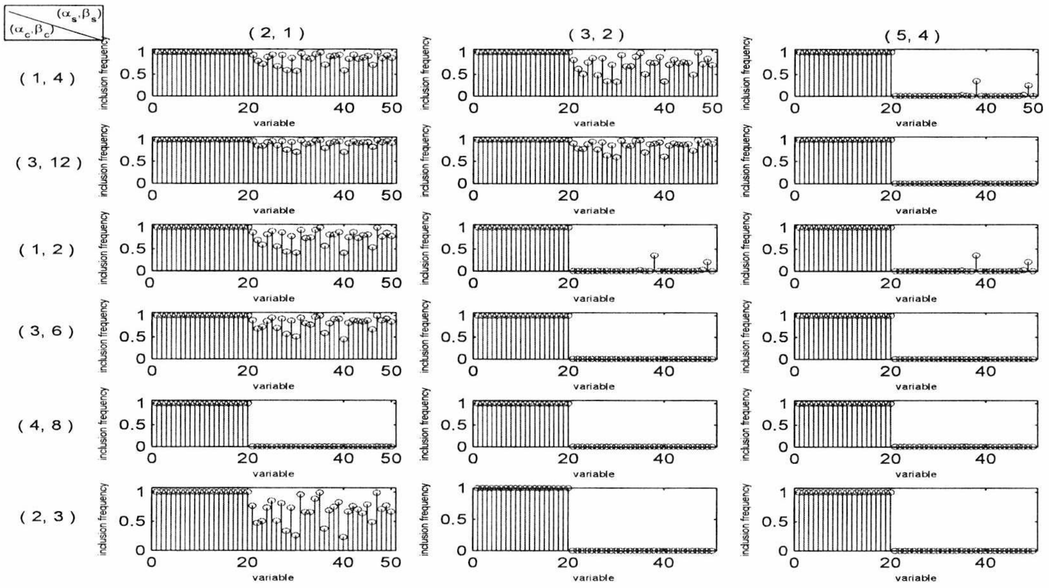

Figure 3.3.5: Cluster-Variable Heterogeneity for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various pairs of (α_c, β_c), (α_s, β_s), assuming heterogeneity and setting $h = 1000$.

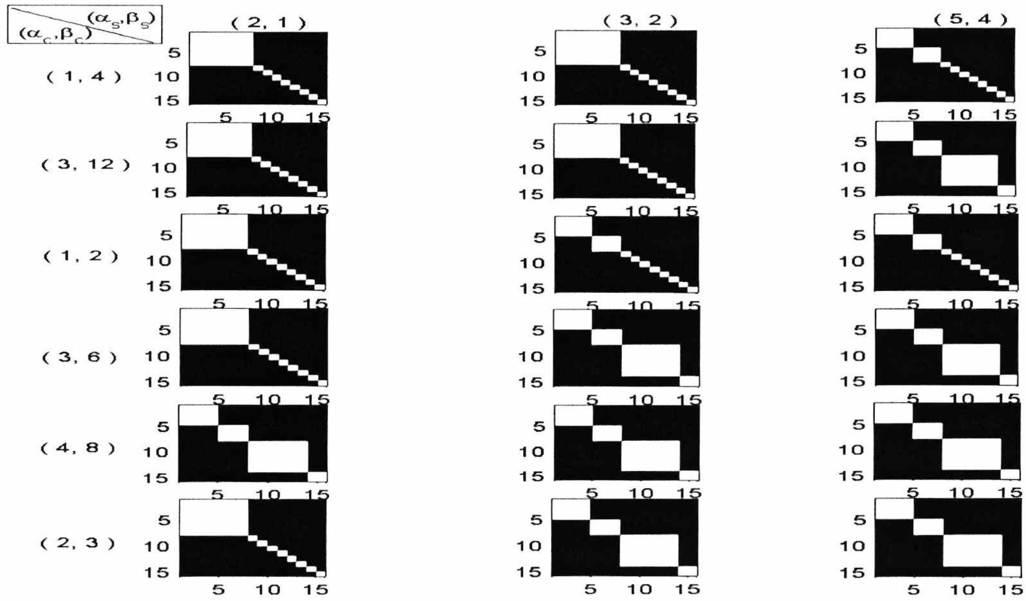

Figure 3.3.6: Cluster-Variable Heterogeneity for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various pairs of (α_c, β_c), (α_s, β_s), assuming heterogeneity and setting $h = 1000$.

To summarise the results on the simulated data set, it looks as though the model is robust with respect to the (α_c, β_c) hyperparameters, as different pairs derive similar results. The hyperparameters h_1, h_0, however, are of major importance, with the (α_s, β_s) hyperparameters playing an important part, being able to drag the chain into achieving convergence, when a bad choice of h_1, h_0 has been made, e.g $h_0 = h_1 = 1000$. The increased amount of hyperparameters though, as we see in the Figures above, results to the need of examining many different combinations of hyperparameters for one to capture possible patterns of bad behaviour of the model. Therefore, the possibility of eliminating the number of hyperparameters to overcome the load of the many different combinations has been under consideration and developed in the following section.

3.4 Cluster-Variable Heterogeneity with a prior on the mean of the $c_k \sim Ga(\alpha_c, \beta_c)$

Motivated by the need of a fewer amount of hyperparameters to be tuned, we considered allowing the algorithm to explore the set of different possible values for the $Ga(\alpha_c, \beta_c)$ prior assigned to the c_k scalars. The idea is while controlling the shape hyperparameter α_c of the Cluster-Variable Heterogeneity model, together with the (α_s, β_s) hyperparameters, we allow m to be the mean of the Gamma prior on c_k's $(m = \alpha_c/\beta_c)$ and we express the joint posteriors in terms of α_c and m by setting $\beta_c = \alpha_c/m$.

For the homogeneous case, we remind that we have chosen to integrate the precisions σ_j^{-2} out, resulting to a joint posterior of the parameters (y, γ, w, G, c) and of form as defined in (3.3.1). Therefore, we maintain the methods described in section 3.3.2 for the sampling of the (γ, y, w), with full conditionals given by

(3.3.2), (3.3.2) and a *Dirichlet* $(\alpha_w + n_1, \ldots, \alpha_w + n_G)$ respectively. However, before performing the random walk search for the scalar c with full conditional as defined in (3.3.4), we allow a second random walk chain to draw samples for $m = \alpha_c/\beta_c$. With a uniform prior on $[0, 1]$ for m and proposing values from a log-normal, we accept m with full conditional :

$$f\left(m|G, w, \gamma, y, X, c\right) \propto p\left(c\right), \tag{3.4.1}$$

where, $p\left(c\right)$ is the Gamma prior set on c. Finally, the Reversible Jump sampler concludes the MCMC chain with the split/merge and birth/death moves being accepted with probability $\min\left(1, A\right)$ and ratios A as given in (3.3.5) and (2.4.9), (2.4.10).

Now, under the assumption of unequal covariances, we follow again the methodology of the Cluster-Variable Heterogeneity model, with the addition of an extra step in the algorithm, the sampling of m. While, the Metropolis search, the Gibbs sampler and the sub-Gibbs strategy, with the full conditionals given in (3.3.7) and (3.3.2) and a *Dirichlet* $(\alpha_w + n_1, \ldots, \alpha_w + n_G)$ are introduced for the sampling of γ, y and w, we allow samples of m to be drawn via a random walk. Likewise the homogeneous case, we use a uniform prior on $[0, 1]$ for m, proposing values from a log-normal. However, this time we need to discuss further the full conditional of m. Although, the c_k scalars have been integrated out, recall we have the joint posterior $f\left(X, y|G, w, \gamma, \sigma_j^{-2}\right)$, we choose to form the full conditional of m conditioned on c_k's. Therefore, we take :

$$f\left(m|G, w, \gamma, y, X, \sigma_j^{-2}, c_k\right) \propto \prod_{k=1}^{G} p\left(c_k\right), \tag{3.4.2}$$

where, again $p\left(c_k\right)$ is the Gamma prior of the k^{th} scalar. Having updated the mean of the Gamma prior of the c_k's, the algorithm continues with the Collapsed-Gibbs sampler for the precisions σ_j^{-2}. Drawing samples of c_k's from the Generalised Inverse Gaussian of form (3.3.14), σ_j^{-2}'s are being generated from Gamma posteriors with parameters as defined in (3.3.11) and (3.3.12). On the final step we have the Reversible Jump technique with the number of components being updated through split/merge and birth/death moves with ratios given in (3.3.15), (2.4.9) and (2.4.10) forming the acceptance probabilities $\min(1, A)$.

3.4.1 Example

Continuing on the application of the simulated data set, assuming heterogeneity we now try the three pairs of (α_s, β_s) [(2, 1), (3, 2) and (5, 4)], with four different values for the hyperparameter α_c, that is 1, 2, 3 and 4. Figures 3.4.1 - 3.4.6 show the histograms of the number of selected variables, the posterior inclusion probabilities and the cluster matrices for the cases of $h_1 = h_0 = 100$ and $h_1 = h_0 = 1000$.

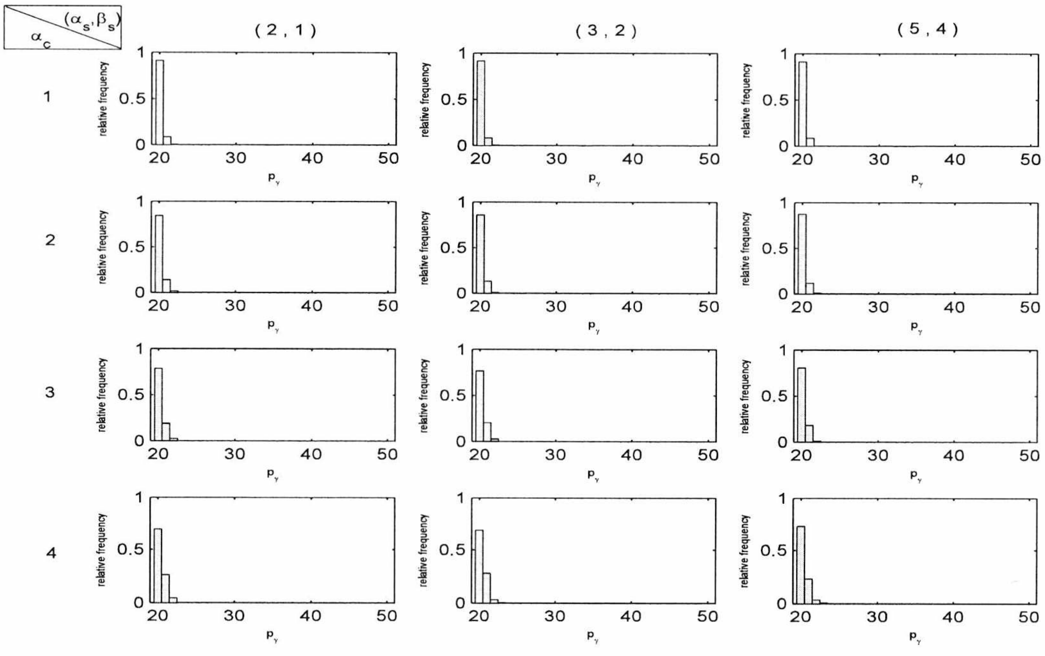

Figure 3.4.1: Cluster-Variable Heterogeneity with a prior on m for the simulated data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 100$.

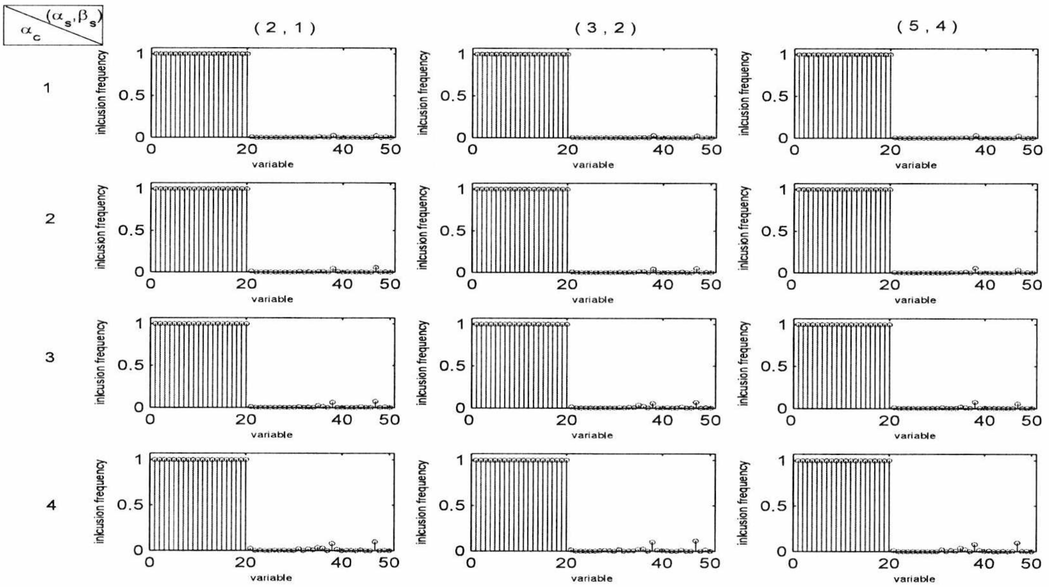

Figure 3.4.2: Cluster-Variable Heterogeneity with a prior on m for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 100$.

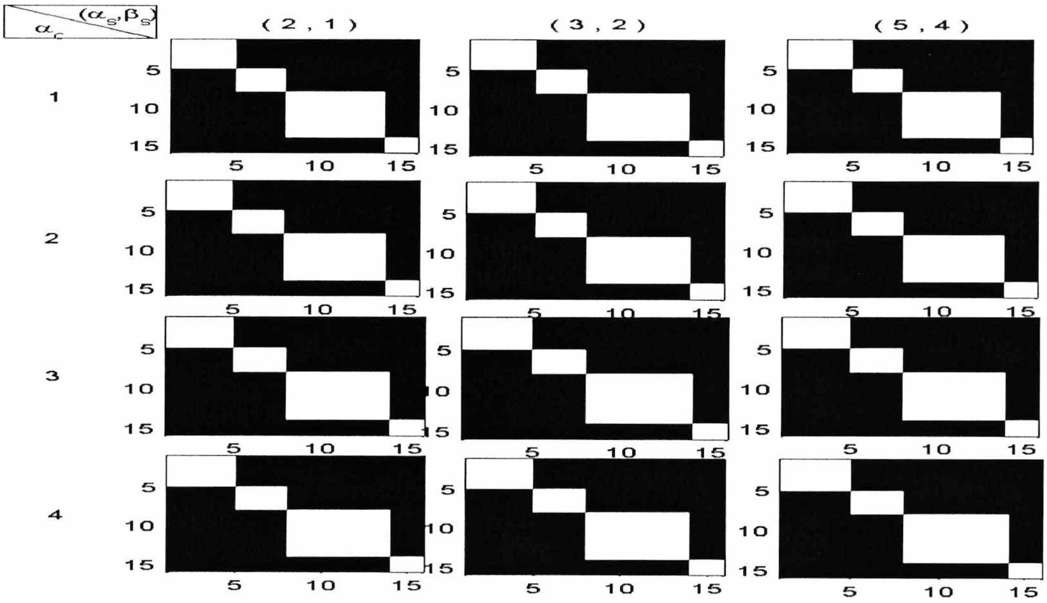

Figure 3.4.3: Cluster-Variable Heterogeneity with a prior on m for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various combinations of (α_s, β_s) and α_c assuming heterogeneity and setting $h = 100$.

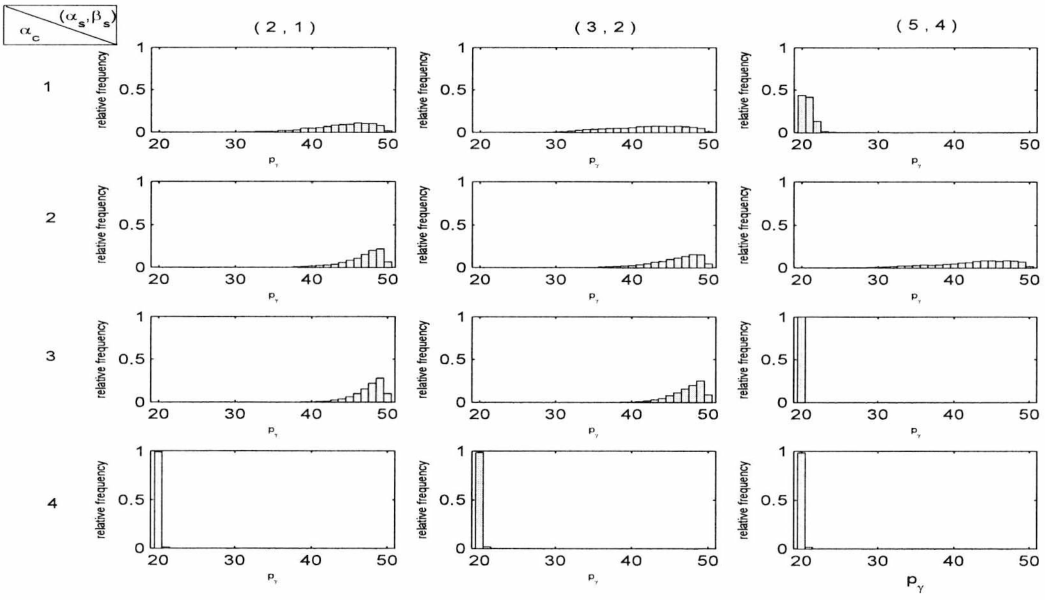

Figure 3.4.4: Cluster-Variable Heterogeneity with a prior on m for the simulated data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 1000$.

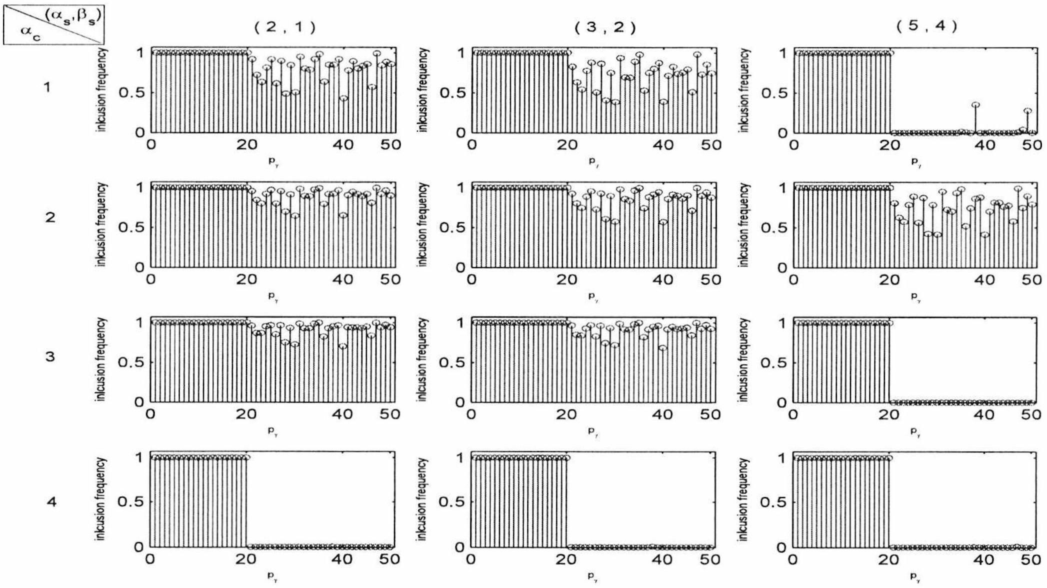

Figure 3.4.5: Cluster-Variable Heterogeneity with a prior on m for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 1000$.

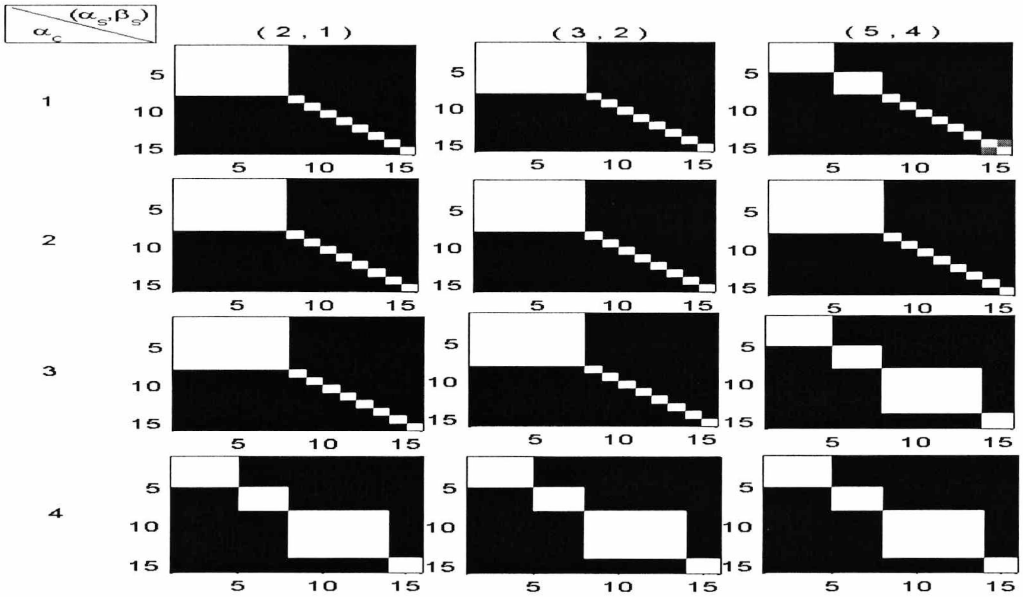

Figure 3.4.6: Cluster-Variable Heterogeneity with a prior on m for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various combinations of (α_s, β_s) and α_c assuming heterogeneity and setting $h = 1000$.

A quick look at the Figures above is enough to let us draw once again the conclusion that the set of h_1 and h_0 is of critical importance and a bad choice can diminish the efficiency of the algorithm in uncovering the cluster structure of the data and simultaneously identifying the important variables. Clearly, under the case of $h_1 = h_0 = 100$, the algorithm achieves convergence to the 20 discriminating variables, recovering the four groups of observations, for all combinations of (α_s, β_s) and α_s (Figures 3.4.1, 3.4.3). However, for $h_1 = h_0 = 1000$, only a few cases (mainly that of $\alpha_c = 4$) can identify the discriminating variables concluding to the formulation of the four groups as expected (Figures 3.4.4, 3.4.6). That is a very valuable remark as it shows that with this new covariance structure for the model, we have managed to stabilise the performance of the algorithm, confining the hyperparameters needing additional care only to the hyperparameters h_1 and h_0. Additionally, as we saw in 3.2.3, setting h_1 equal to h_0 is absolutely reasonable; therefore, instead of h_1 and h_0, we can claim that the crucial hyperparameter is only one and that is, say, h.

3.5 Cluster-Variable Heterogeneity with a prior on the mean of the $c_k \sim Ga\left(\alpha_c, \beta_c\right)$ and a prior on hyperparameter h

Having come to the conclusion that the algorithm has a sensitivity on the value of the hyperparameter h and that a "bad" choice can actually be detrimental to the recovery of the cluster structure and the identification of the discriminating variables, means to tackle this sensitivity out are of interest. Naturally, the introduction of a hyperprior on h is the direction that we followed. Therefore, for the Cluster-Variable Heterogeneity model with a prior on m, we built on the prior specifications by introducing an Inverse Gamma prior on h, i.e.

$$h \sim IG\left(\alpha_h, \beta_h\right). \tag{3.5.1}$$

Starting with the homogeneous case and building up the algorithm of the Cluster-Variable Heterogeneity model with a prior on m, we initialise the runs with the sampling of h using a random walk chain, proposing moves from a log-normal and accepting with full conditional

$$f\left(h|G, w, \gamma, y, X, c\right) \propto f\left(X, y|G, w, \gamma, c, h\right) p\left(h\right). \tag{3.5.2}$$

The algorithm continues with the sampling of (γ, y, w, G, c) together with m according to our description for the Cluster-Variable Heterogeneity model with the prior on m.

Finally, assuming heterogeneity, we sample h, similarly to the case of homogeneous covariances using a random walk, but this time the full conditional of h

is :

$$f\left(h|G,w,\gamma,y,X,\sigma_j^{-2}\right) \propto f\left(X,y|G,w,\gamma,\sigma_j^{-2},h\right)p\left(h\right). \qquad (3.5.3)$$

for which, we note that unlike m, we do not need to condition on c_k's. The sampling of the remaining $\left(\gamma,y,w,G,\sigma_j^{-2}\right)$ parameters, as this has been described earlier for the CVH model with a prior on m, will complete the chain.

3.5.1 Example

We will now close this set of the different covariance structures with the application of the latter idea of considering an additional hyperprior on h and we will examine whether we have managed to overcome the sensitivity imposed by its crucial choice.

Following the settings of the example under Section 3.4.1, we examined the behaviour of the model for two sets of values for the hyperparameters of the $IG\left(\alpha_h,\beta_h\right)$ prior of h. For pairs of (α_h,β_h), $(3,200)$ and $(3,400)$, additionally to the usual histograms for the number of selected variables (Figures 3.5.1, 3.5.5), the posterior inclusion probabilities (Figures 3.5.2, 3.5.6) and the matrices indicating the cluster structure as suggested by the chains (Figures 3.5.3, 3.5.7), this time we also display the posterior distributions of h signifying its estimated value (Figures 3.5.4, 3.5.8).

The results are rather pleasing as for all cases, no matter the settings of the hyperparameters of h, we achieved convergence on the recovery of the discriminating variables as well as on the construction of the four components as they had been originally designed. Of course, we should not forget to notice that for the case of $(3,400)$ and with a $\alpha_c=4$, we have a slower convergence of p_γ, but the algorithm does eventually identify the 20 important variables.

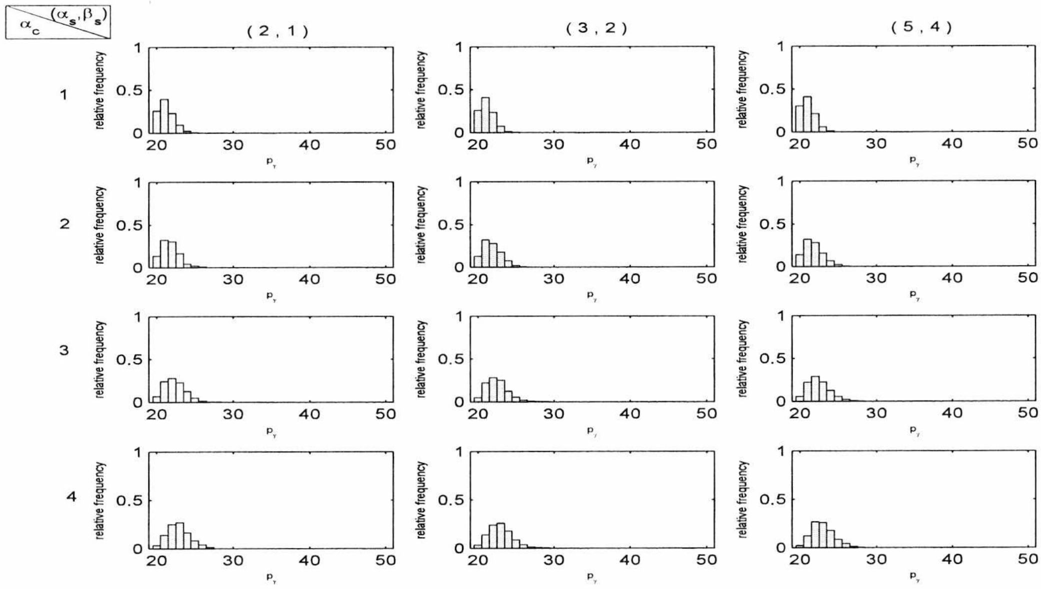

Figure 3.5.1: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

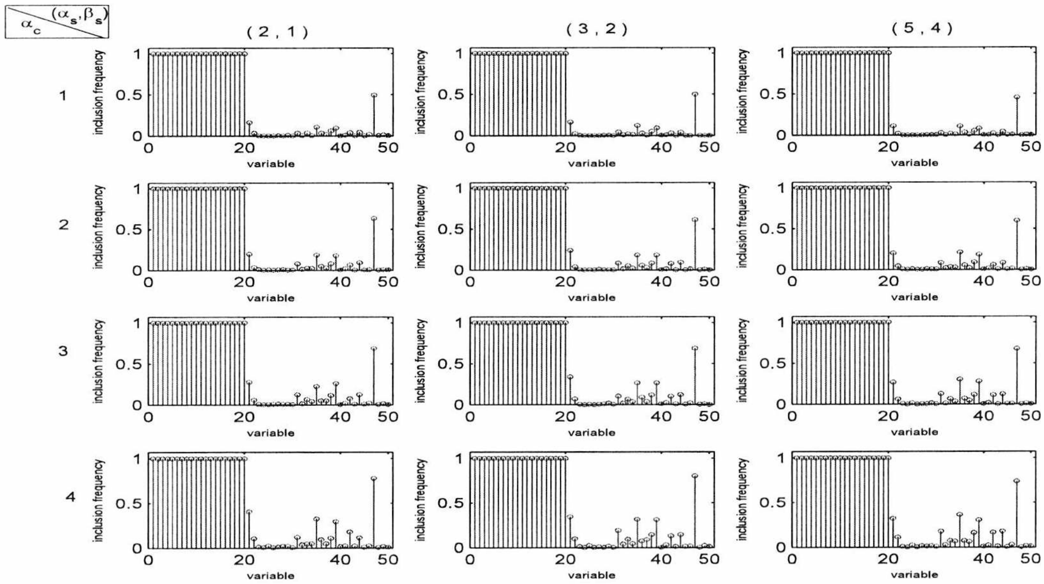

Figure 3.5.2: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

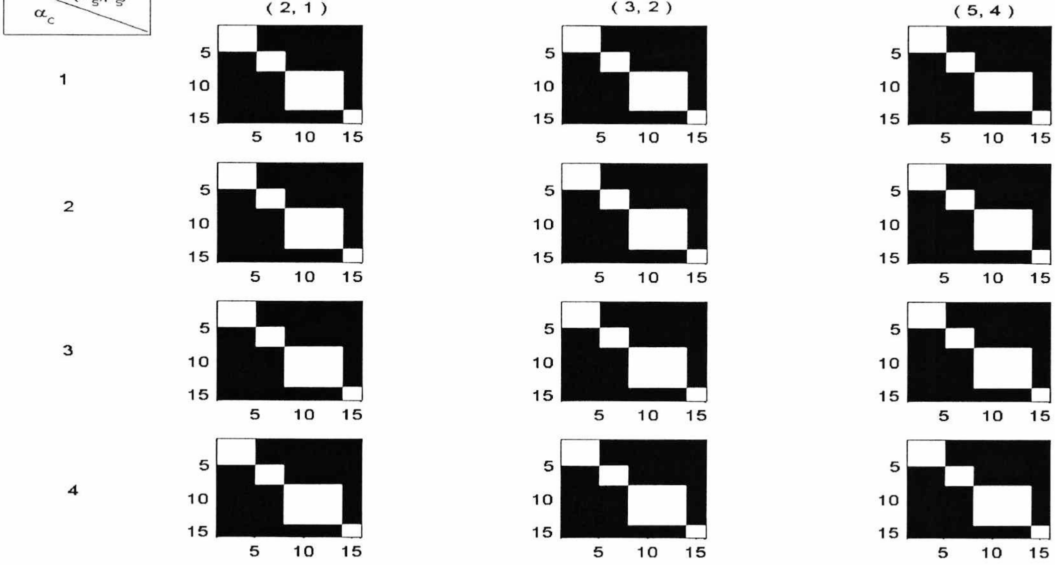

Figure 3.5.3: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

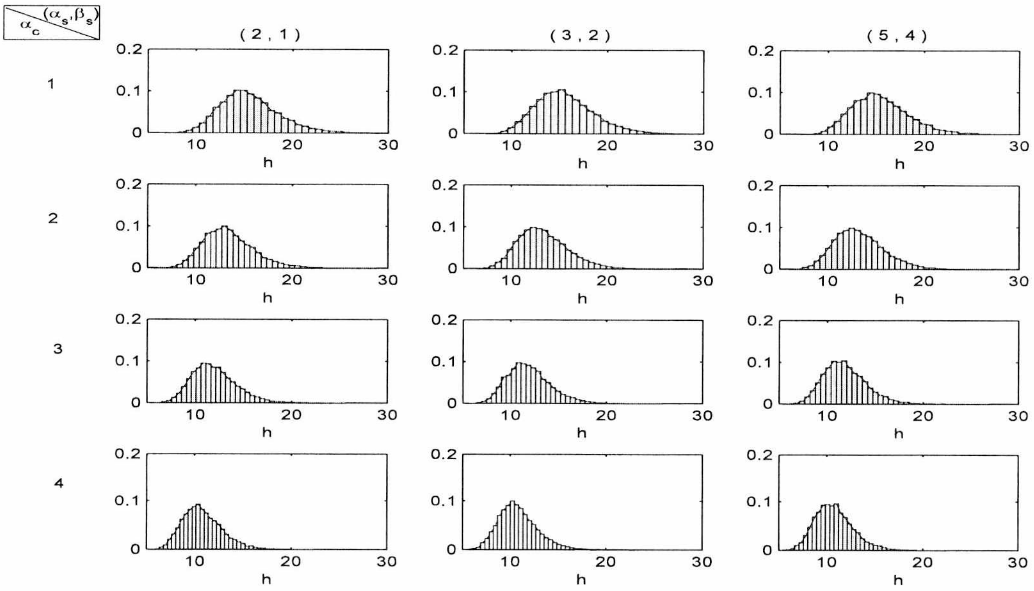

Figure 3.5.4: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

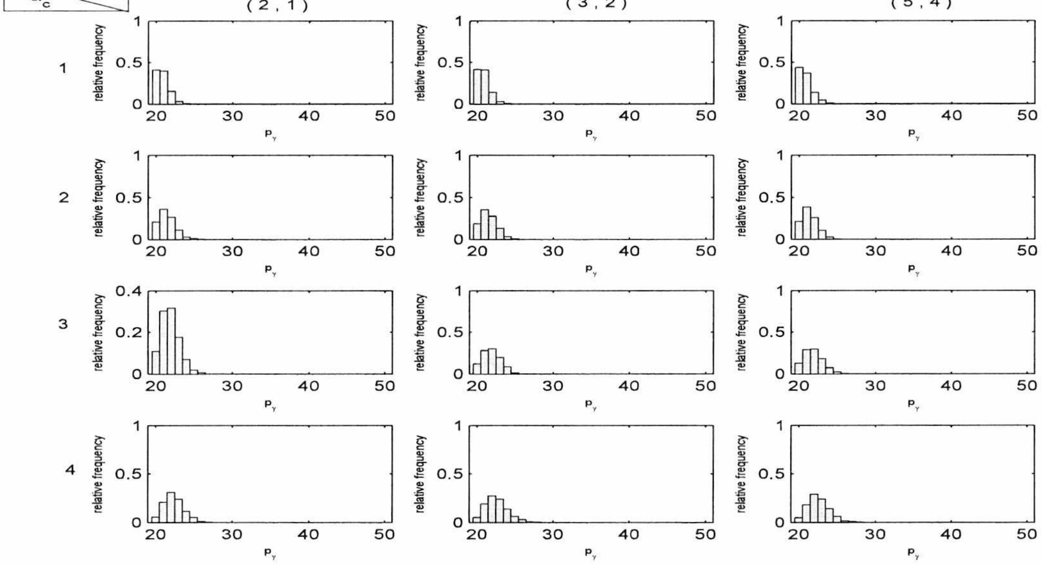

Figure 3.5.5: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

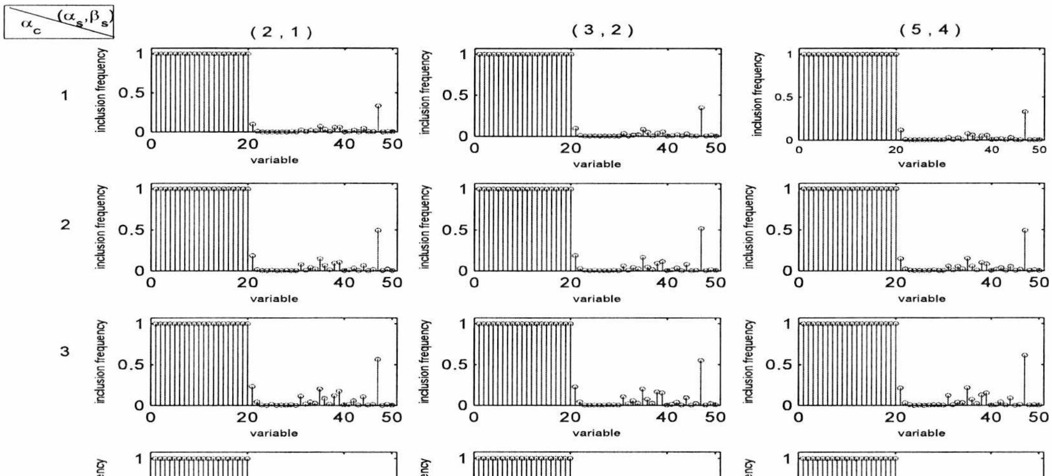

Figure 3.5.6: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

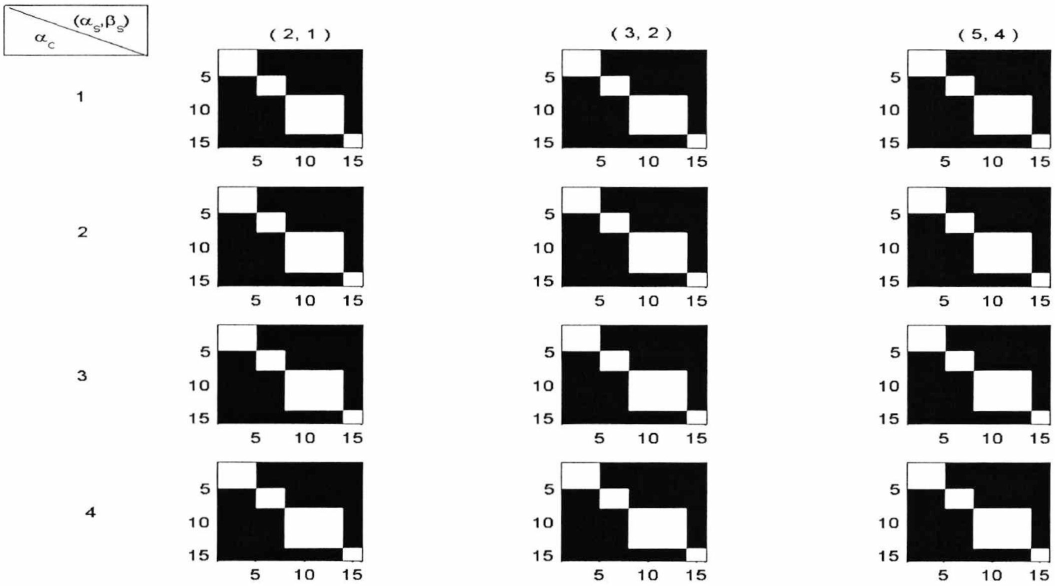

Figure 3.5.7: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 400)$, with unequal covariance matrices.

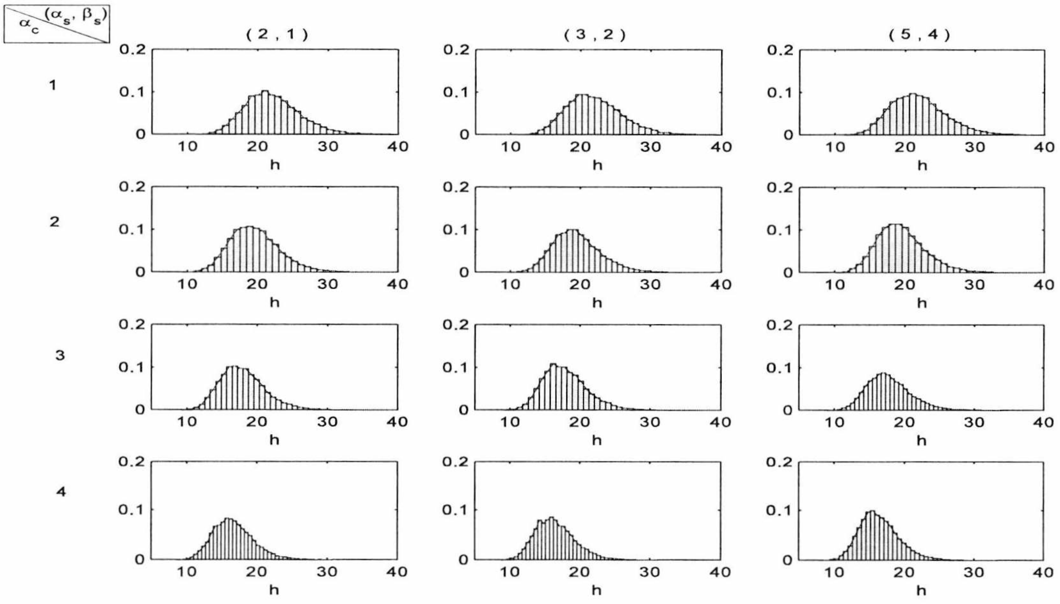

Figure 3.5.8: Cluster-Variable Heterogeneity with priors on m and h for the simulated data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 400)$, with unequal covariance matrices.

Finally, looking at the posterior distributions of h in Figures 3.5.4 and 3.5.8, one can observe something very interesting. The estimated value of h varies between 10 and 12; case for which we did not manage to achieve convergence for the cluster assignment under the CH model (Figures 3.2.2, 3.2.6), while for the covariance structure of the CVH model an h of value 100 produced equally good results (Figure 3.3.3).

For all models we have examined so far we have observed the great importance of the choice of the h hyperaparameter. The different values on h that we have previously experimented with $(1, 10, 100, 1000)$ usually led to variant cluster structures and sets of discriminating variables, lacking criteria on which value of h supported convergence to the expected set of the 20 variables uncovering the structure of four clustres. In any case, we can now say that setting a prior on h has facilitated the convergence of the algorithm. We have therefore managed to overcome the model's sensitivity on h building up a robust model.

3.6 Non-Conjugate Covariance Structure

So far, in the model of chapter 2, as well as in the different covariance structures presented earlier in this chapter, we have repeatedly seen the importance of the hyperparameter h of the Inverse-Wishart prior assigned to the covariance matrices Σ_k^D and Σ^{ND}. Setting a hyperprior on h, we managed to bypass the need of choosing a proper value for it, however, adopting a model where h is not considered may be necessary. The importance of h lies on the fact that it consists of a linkage of the within-cluster variation and the between-cluster variation. Consequently, in problems with clusters of different magnitude for instance, assuming a single - and common for all clusters - scalar h, could affect the recovery of the true cluster structure. That is, having a large value of h, would most likely capture the big clusters, however, that would also result in the mean values of the smaller groups having a big variance. Therefore, smaller groups with small between cluster variation could be captured under a single big normal distribution, rather than many smaller ones.

We understand, thus, the need of regarding the model with a cluster structure of the non-conjugate form. For

$$x_i^D | y_i = k, \theta, \gamma \sim N\left(\mu_k^D, \Sigma_k^D\right), \quad x_i^{ND} | \theta^*, \gamma \sim N\left(\mu^{ND}, \Sigma^{ND}\right),$$

and assuming covariances matrices

$$\Sigma^D = cB_{p_\gamma}, \quad \Sigma^{ND} = B_{p-p_\gamma},$$

under the case of homogeneous covariances and

$$\Sigma_k^D = c_k B_{p_\gamma}, \quad \Sigma^{ND} = B_{p-p_\gamma},$$

when under heterogeneity, we will examine the non-conjugate model. As previously :

$$B = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \sigma_p^2 \end{pmatrix}.$$

3.6.1 Prior Settings

Starting with the prior on the number of discriminating variables p_γ, we set a beta-binomial distribution with parameters a, b such that $a + b = 2$. The number of components G follows a priori a discrete uniform on $[1, \ldots, G_{max}]$, while the prior distribution of the component weights w is a $Dirichlet\,(\alpha_w, \ldots, \alpha_w)$. Using the prior information, we assign normal prior distributions on the mean vectors μ_k^D, μ^{ND}, such that :

$$\begin{aligned} \mu_k^D &\sim N\left(\mu_0^D, \Sigma_0^D\right), \\ \mu^{ND} &\sim N\left(\mu_0^{ND}, \Sigma_0^{ND}\right), \end{aligned}$$

where, Σ_0^D and Σ_0^{ND} are submatrices of the diagonal matrix of dimension p, Σ_0, which in turn is an estimate of the data. We set Σ_0 equal to a small multiple, k, of a diagonal matrix, say $\widehat{\Sigma}$, i.e.

$$\Sigma_0 = k\widehat{\Sigma},$$

for which, we have :

$$\widehat{\Sigma} = \frac{1}{1.349^2}\begin{pmatrix} r_1^2 & 0 & \ldots & 0 \\ 0 & r_2^2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \ldots & 0 & r_p^2 \end{pmatrix},$$

with $r_j, j = 1, \ldots, p$ being the interquartile range of the p variables (Cramer, 1946).

The precisions σ_j^{-2}'s, $j = 1, \ldots, p$, for both cases of homogeneity and heterogeneity, follow a priori a Gamma distribution with parameters (α_s, β_s),

$$\sigma_j^{-2} \sim Ga\left(\alpha_s, \beta_s\right),$$

while, for the scalars c_k's, $k = 1, \ldots, G$, as well as the single scalar c, we have Gamma priors with parameters (α_c, β_c), i.e.

$$c_k \sim Ga\left(\alpha_c, \beta_c\right), \quad \text{and} \quad c \sim Ga\left(\alpha_c, \beta_c\right).$$

3.6.2 Posterior Inference

Moving to the posterior formulation of the model and examining the case of assuming unequal covariances, we choose to integrate the mean vectors μ_k^D, μ^{ND} out, obtaining the joint posterior for the parameters $\left(\gamma, y, w, G, \sigma_j^{-2}, c_k\right)$. That is :

$$\begin{aligned}
& f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c_k\right) = (2\pi)^{-\frac{np}{2}} \left|\Sigma_0^D\right|^{-\frac{G}{2}} \left|\Sigma_0^{ND}\right|^{-\frac{1}{2}} \left|B_{p-p_\gamma}\right|^{-\frac{n}{2}} \left|\Sigma_0^{ND^{-1}} + nB_{p-p_\gamma}^{-1}\right|^{-\frac{1}{2}} \\
& \times \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} \left(x_i^{ND} - \overline{x}^{ND}\right)' B_{p-p_\gamma}^{-1} \left(x_i^{ND} - \overline{x}^{ND}\right) \right. \\
& \left. + n\left(\mu_0^{ND} - \overline{x}^{ND}\right)' \Sigma_0^{ND^{-1}} \left(\Sigma_0^{ND^{-1}} + nB_{p-p_\gamma}^{-1}\right)^{-1} B_{p-p_\gamma}^{-1} \left(\mu_0^{ND} - \overline{x}^{ND}\right) \right\} \\
& \times \prod_{k=1}^{G} \left\{ w_k^{n_k} c_k^{-\frac{n_k p_\gamma}{2}} \left|B_{p_\gamma}\right|^{-\frac{n_k}{2}} \left|\Sigma_0^{D^{-1}} + n_k c_k^{-1} B_{p_\gamma}^{-1}\right|^{-\frac{1}{2}} \right. \\
& \times \exp\left[-\frac{1}{2} \sum_{x_i \in C_k} \left(x_i^D - \overline{x}_k^D\right)' c_k^{-1} B_{p_\gamma}^{-1} \left(x_i^D - \overline{x}_k^D\right) \right. \\
& \left.\left. + n_k \left(\mu_0^D - \overline{x}_k^D\right)' \Sigma_0^{D^{-1}} \left(\Sigma_0^{D^{-1}} + n_k c_k^{-1} B_{p_\gamma}^{-1}\right)^{-1} c_k^{-1} B_{p_\gamma}^{-1} \left(\mu_0^D - \overline{x}_k^D\right) \right] \right\},
\end{aligned} \qquad (3.6.1)$$

with, $\overline{x}_k^D$ being the sample mean of the k^{th} cluster for the covariates that are included in the model and $\overline{x}^{ND}$ the overall sample mean of the non-discriminating variables.

Previously in this chapter, and more specifically for the variable heterogeneity model (section 3.3), we adopted the marginalisation of the joint posterior over the scalars c_k's. However, under the non-conjugate case, the integration of the joint posterior in (3.6.1) is not trivial for neither the variances σ_j's nor scalars c_k's. Therefore, assuming heterogeneous covariances, we choose to sample all $\left(\gamma, w, G, y, \sigma_j^{-2}, c_k\right)$.

Starting with the sampling of the latent vector γ and using a Metropolis search, an Add, Delete or Swap move randomly proposes a new γ' candidate. With full conditional :

$$f\left(\gamma|G, w, y, X, \sigma_j^{-2}, c_k\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c_k\right) p\left(\gamma|G\right), \qquad (3.6.2)$$

we accept the proposed vector with probability $\min\left[1, \frac{f(\gamma'|X,y,w,G)}{f(\gamma|X,y,w,G)}\right]$.

On the following, with the component weights w being drawn from a Dirichlet distribution with parameters $(\alpha_w + n_1, \ldots, \alpha_w + n_G)$ via a Gibbs sampler, a sub - Gibbs strategy is used to update the elements of the allocation vector y one at a time, with a full conditional conditioned on the preciously allocated observations, i.e.

$$f\left(y|G, w, \gamma, X, \sigma_j^{-2}, c_k\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c_k\right). \qquad (3.6.3)$$

Next, we need to sample scalars c_k's and the precisions σ_j^{-2}. Using the idea of the Collapsed-Gibbs sampler, we sample both c_k's, σ_j^{-2}, blocking on the mean vectors μ_k^D's, μ^{ND}, although we had originally integrated them out. For Normal full conditionals such that :

$$\mu_k^D \sim N\left(\mu_0^{D*}, \Lambda_1^D\right), \quad \mu^{ND} \sim N\left(\mu_0^{ND*}, \Lambda_2^{ND}\right), \qquad (3.6.4)$$

where,

$$\mu_0^{D*} = \left(\Sigma_0^{D^{-1}} + n_k \Sigma_k^{D^{-1}}\right)^{-1} \left(\Sigma_0^{D^{-1}} \mu_0^D + n_k \Sigma_k^{D^{-1}} \overline{x}_k^D\right),$$
$$\mu_0^{ND*} = \left(\Sigma_0^{ND^{-1}} + n\Sigma^{ND^{-1}}\right)^{-1} \left(\Sigma_0^{ND^{-1}} \mu_0^{ND} + n\Sigma^{ND^{-1}} \overline{x}^{ND}\right),$$

$$\Lambda_1^D = \left(\Sigma_0^{D^{-1}} + n_k \Sigma_k^{D^{-1}}\right)^{-1}, \quad \Lambda_2^{ND} = \left(\Sigma_0^{ND^{-1}} + n \Sigma^{ND^{-1}}\right)^{-1},$$

Gibbs samplers are being used to draw samples for the mean vectors. Conditioned on the values of μ_k^D, μ^{ND}, the resulting full conditionals of c_k, σ_j^{-2} are of closed form and therefore Gibbs samplers can also be applied. In particular, the full conditional of the c_k scalars is a Generalised Inverse Gaussian distribution,

$$c_k \sim GIG\left(p_c, a_c, b_c\right),$$

with parameters $b_c = \sum\limits_{x_i \in k} \left(x_i^D - \mu_k^D\right)' B_{p_\gamma}^{-1} \left(x_i^D - \mu_k^D\right)$, $p_c = \alpha_c - n_k p_\gamma / 2$ and $a_c = 2\beta_c$.

On the other hand, for the precisions σ_j^{-2}, $j = 1, \ldots, p$, we distinguish between those associated to the discriminating variables and the corresponding to the non-discriminating ones. Both sets of precisions follow a posteriori, and always conditioned on $\left(\mu_k^D, \mu^{ND}\right)$, Gamma distributions of the form :

$$\sigma_j^{-2} \sim Ga\left(\alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2} \sum_{k=1}^{G} \sum_{x_i \in k} \left(x_{ij}^D - \mu_{kj}^D\right)' \frac{1}{c_k} \left(x_{ij}^D - \mu_{kj}^D\right)\right), \tag{3.6.5}$$

for the discriminating variables and

$$\sigma_{j(\gamma^c)}^{-2} \sim Ga\left(\alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2} \sum_{i=1}^{n} \left(x_{ij}^{ND} - \mu_j^{ND}\right)' \left(x_{ij}^{ND} - \mu_j^{ND}\right)\right), \tag{3.6.6}$$

for the non-discriminating ones.

Coming now to the Reversible Jump sampler, evidently, alike the component weights, this time the dimension of the c_k scalars also alters as we jump between different states. Choosing randomly between a split and a merge move, with probabilities b_G and d_G, let us first consider a split move. Say an empty or non-empty component ℓ, is chosen to be split, obtaining the clusters ℓ_1, ℓ_2. The new clusters are assigned the weights $w'_{\ell_1} = w_\ell u$ and $w'_{\ell_2} = w_\ell (1-u)$, with $u \sim Be\,(2,2)$. However, new scalars c'_{ℓ_1} and c'_{ℓ_2} need also being considered. We choose to draw values for c'_{ℓ_1} and c'_{ℓ_2} from the $Ga\,(\alpha_c, \beta_c)$ prior set on the c_k's.

Being in state $\psi = \left(G, w, \gamma, \sigma_j^{-2}, y, c\right)$, split move will take us to state $\psi' = \left(G+1, w', \gamma, \sigma_j^{-2}, y', c'\right)$, where c and c' are regarded as the vectors containing the c_k scalars for $k = 1, \ldots, G$ and $k = 1, \ldots, G+1$ respectively. Having suggested the new scalars and component weights, the move is being accepted with probability $\min\,(1, A)$, where for :

$$A = \frac{p\,(\psi'|x)\, r_m\,(\psi')\, p\,(c)}{p\,(\psi|x)\, r_m\,(\psi)\, p\,(c')\, q\,(u)} \left| \frac{\partial(w'_{\ell_1}, w'_{\ell_2})}{\partial\,(w_\ell, u)} \right|, \tag{3.6.7}$$

we define the terms :

$$\begin{aligned} p\,(\psi'|x) &= f\left(G+1, w', \gamma, \sigma_j^{-2}, y', c'|X\right) \\ &= f\left(X, y'|w', \gamma, \sigma_j^{-2}, G+1, c'\right) f\,(w'|G+1)\, f\,(G+1)\, p\,(c'), \\ p\,(\psi|x) &= f\left(G, w, \gamma, \sigma_j^{-2}, y, c|X\right) \\ &= f\left(X, y|w, \gamma, \sigma_j^{-2}, G, c\right) f\,(w|G)\, f\,(G)\, p\,(c), \\ r_m\,(\psi) &= \frac{b_G p_{alloc}}{G}, \\ r_m\,(\psi') &= \frac{d_{G+1}}{G(G+1)}, \\ q\,(u) &= \frac{1}{B\,(2,2)} u\,(1-u), \qquad \left| \frac{\partial\left(w'_{\ell_1}, w'_{\ell_2}\right)}{\partial\,(w_\ell, u)} \right| = w_\ell, \end{aligned}$$

with $p_{alloc} = u^{n'_{\ell_1}}\,(1-u)^{n'_{\ell_2}}$ the probability that the particular allocation has been made. Replacing in (3.6.7), we obtain :

$$A = \frac{f\,(X, y'|w', \gamma, G+1, c')}{f\,(X, y|w, \gamma, G, c)} \times \frac{d_{G+1}}{b_G} \times \frac{B\,(2,2)}{B\,(\alpha_w, G\alpha_w)} \times u^{\alpha_w - 1} \times (1-u)^{\alpha_w - 1} \times w_\ell^{\alpha_w}, \tag{3.6.8}$$

where, $f\left(X, y|w', \gamma, G+1, c'\right)$ and $f\left(X, y|w, \gamma, G+1, c\right)$ as derived in (3.6.1).

Proposing the reverse move, i.e. two components are randomly chosen to be merged, we jump from state $\psi = \left(G+1, w, \gamma, \sigma_j^{-2}, y, c\right)$ to state $\psi' = (G, w', \gamma, \sigma_j^{-2}, y, c')$, drawing the variable $u = w_{\ell_1}/w'_{\ell}$ from a $Be\left(2, 2\right)$. The weight of the merged component is now $w'_{\ell} = w_{\ell_1} + w_{\ell_2}$, while its scalar c'_{ℓ} is being sampled from the $Ga\left(\alpha_c, \beta_c\right)$ prior. We accept the merge move with probability $\min\left(1, A\right)$, where A is the reverse ratio obtained for the split move (3.6.8).

Finally, we conclude the RJMCMC sampler of the non-conjugate model for the heterogeneous case, with the generation or deletion of an empty component. In the case of a birth move, the weight of the new empty component w'_{G+1} is generated from a $Be(1, G)$, with the weights of the remaining components being rescaled such that $w'_k = w_k\left(1 - w'_{G+1}\right)$; a strategy followed in all the birth moves we have developed so far. However, the additional c'_{G+1} scalar, corresponding to the new component, is drawn, similarly to the split/merge moves, from the Gamma prior, of the c_k's, $Ga\left(\alpha_c, \beta_c\right)$, while unlike w_k's, the remaining G scalars do not need to be rescaled. The birth move is finally being accepted with probability $\min\left(1, A\right)$, for which, ratio A is given by:

$$A = \left(1 - w'_{G+1}\right)^n \times \frac{w'^{(\alpha_w-1)}_{G+1}\left(1 - w'_{G+1}\right)^{G(\alpha_w-1)}}{B\left(\alpha_w, G\alpha_w\right)} \times \frac{d_{G_0+1}}{\left(G_0+1\right) b_{G_0}} \times \frac{G+1}{G}. \tag{3.6.9}$$

Deleting an empty component, say w_{G+1}, we also need to delete its corresponding scalar c_{G+1} and rescale the weights according to $w'_k = w_k/(1 - w_{G+1})$, for $k = 1, \ldots, G-1$. Ratio A of the acceptance probability $\min\left(1, A\right)$ now becomes:

$$A = \left(1 - w_{G+1}\right)^{-n} \times \frac{B\left(\alpha_w, G\alpha_w\right)}{w^{(\alpha_w-1)}_{G+1}\left(1 - w_{G+1}\right)^{G(\alpha_w-1)}} \times \frac{\left(G_0+1\right) b_{G_0}}{d_{G_0+1}} \times \frac{G}{G+1}. \tag{3.6.10}$$

For the end, we shall discuss the settings for the case of homogeneous covariances. Marginalising, as in the case of heterogeneity, over the mean vectors μ_k^D, μ^{ND}, the parameters of interest now become : $\left(\gamma, y, w, G, \sigma_j^{-2}, c\right)$, the joint posterior of which is :

$$\begin{aligned}
f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c\right) &= (2\pi)^{-\frac{np}{2}} \left|\Sigma_0^D\right|^{-\frac{G}{2}} \left|\Sigma_0^{ND}\right|^{-\frac{1}{2}} \left|B_{p-p_\gamma}\right|^{-\frac{n}{2}} c^{-\frac{np_\gamma}{2}} \left|B_{p_\gamma}\right|^{-\frac{n}{2}} \\
&\times \left|\Sigma_0^{ND^{-1}} + nB_{p-p_\gamma}^{-1}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n} \left(x_i^{ND} - \overline{x}^{ND}\right)' B_{p-p_\gamma}^{-1} \left(x_i^{ND} - \overline{x}^{ND}\right) \right. \\
&\left. + n\left(\mu_0^{ND} - \overline{x}^{ND}\right)' \Sigma_0^{ND^{-1}} \left(\Sigma_0^{ND^{-1}} + nB_{p-p_\gamma}^{-1}\right)^{-1} B_{p-p_\gamma}^{-1} \left(\mu_0^{ND} - \overline{x}^{ND}\right) \right\} \\
&\times \prod_{k=1}^{G} \left\{ w_k^{n_k} \left|\Sigma_0^{D^{-1}} + n_k c^{-1} B_{p_\gamma}^{-1}\right|^{-\frac{1}{2}} \right. \\
&\times \exp\left[-\frac{1}{2} \sum_{x_i \in C_k} \left(x_i^D - \overline{x}_k^D\right)' c^{-1} B_{p_\gamma}^{-1} \left(x_i^D - \overline{x}_k^D\right) \right. \\
&\left.\left. + n_k \left(\mu_0^D - \overline{x}_k^D\right)' \Sigma_0^{D^{-1}} \left(\Sigma_0^{D^{-1}} + n_k c^{-1} B_{p_\gamma}^{-1}\right)^{-1} c^{-1} B_{p_\gamma}^{-1} \left(\mu_0^D - \overline{x}_k^D\right) \right] \right\}. \qquad (3.6.11)
\end{aligned}$$

Using the Add/Delete/Swap moves and the Metropolis search, we update γ with acceptance probability $\min\left[1, \frac{f(\gamma'|X,y,w,G)}{f(\gamma|X,y,w,G)}\right]$, for which, the full conditional of γ is given by :

$$f\left(\gamma|G, w, y, X, \sigma_j^{-2}, c\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c\right) p\left(\gamma|G\right). \qquad (3.6.12)$$

For the component weights, we have the *Dirichlet* $\left(\alpha_w + n_1, \ldots, \alpha_w + n_G\right)$ posterior distribution and a Gibbs sampler drawing samples of w, while with

$$f\left(y|G, w, \gamma, X, \sigma_j^{-2}, c\right) \propto f\left(X, y|G, w, \gamma, \sigma_j^{-2}, c\right), \qquad (3.6.13)$$

and a sub-Gibbs strategy, we sample y on element at a time.

Following the idea of sampling c_k's, σ_j^{-2}'s conditioned on μ_k^D, μ^{ND}, using a Collapsed-Gibbs strategy, we continue with the sampling of the mean vectors. Using Gibbs samplers and Normal distributions as defined in (3.6.4), we draw samples of μ_k^D, μ^{ND} similar to the heterogeneous case. The single scalar c, can now be sampled from a Generalised Inverse Gaussian with parameters $b_c = \sum\limits_{x_i \in k} \left(x_i^D - \mu_k^D\right)' B_{p_\gamma}^{-1} \left(x_i^D - \mu_k^D\right)$, $p_c = \alpha_c - np_\gamma/2$ and $a_c = 2\beta_c$, while the following Gamma distributions :

$$Ga\left(\alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2}\sum_{k=1}^{G}\sum_{x_i \in k}\left(x_{ij}^D - \mu_{kj}^D\right)' \frac{1}{c}\left(x_{ij}^D - \mu_{kj}^D\right)\right), \tag{3.6.14}$$

for the discriminating variables and

$$Ga\left(\alpha_s + \frac{n}{2}, \beta_s + \frac{1}{2}\sum_{i=1}^{n}\left(x_{ij}^{ND} - \mu_j^{ND}\right)'\left(x_{ij}^{ND} - \mu_j^{ND}\right)\right), \tag{3.6.15}$$

for the non-discriminating ones, give us the samples for the precisions σ_j^{-2}.

Examining the Reversible Jump MCMC for the homogeneous case, we go back into the case where weights w are the only parameters need updating to match the change of dimensions. Letting, with probability b_G, the split of a component ℓ into components ℓ_1 and ℓ_2, and with a $u \sim Be(2,2)$, we have the new component weights $w'_{\ell_1} = w_\ell u$ and $w'_{\ell_2} = w_\ell (1-u)$ respectively. From state $\psi = \left(G, w, \gamma, \sigma_j^{-2}, y, c\right)$, we accept the jump to state $\psi' = \left(G+1, w', \gamma, \sigma_j^{-2}, y', c\right)$ with probability $\min(1, A)$, with

$$A = \frac{f\left(X, y'|w', \gamma, G+1, c\right)}{f\left(X, y|w, \gamma, G, c\right)} \times \frac{d_{G+1}}{b_G} \times \frac{B\left(2,2\right)}{B\left(\alpha_w, G\alpha_w\right)} \times u^{\alpha_w - 1} \times (1-u)^{\alpha_w - 1} \times w_\ell^{\alpha_w}, \tag{3.6.16}$$

where, $f\left(X, y|w', \gamma, G+1, c\right)$ and $f\left(X, y|w, \gamma, G+1, c\right)$ are obtained from equation (3.6.11). Reversing the ratio of equation (3.6.16), we accept, with probability

$\min(1, A)$, the merge of components ℓ_1, ℓ_2 into a component, say ℓ, with weight $w'_\ell = w_{\ell_1} + w_{\ell_2}$. This time variable $u = w_{\ell_1}/w'_\ell$ is drawn from a $Be\,(2,2)$.

For the birth/death set of moves, starting with the generation of an empty component with weight w'_{G+1} drawn from a $Be\,(1, G)$, we rescale the weights of the remaining components, i.e. $w'_k = w_k(1 - w'_{G+1})$ and we accept the suggested move with probability $\min(1, A)$, where A is defined as in equation (3.6.9). For the reverse move of the deletion of a component w_{G+1}, we rescale the weights according to : $w'_k = w_k/(1 - w_{G+1})$, for $k = 1, \ldots, G-1$. Again, ratio A of the acceptance probability, $\min(1, A)$, for the death move is given by equation (3.6.10).

3.6.3 Example

Finishing the non-conjugate formulation of the covariance structure of the model, we examine the results from its application on our simulated data set assuming heterogeneity. Iterating for a 60000 times with a burn in of 40000 iterations and with $p_{prior} = 10$, $a + b = 2$, $G_{max} = 15$ and $\alpha_w = 1$, for the shape hyperparameter of the prior on c_k's, i.e. α_c, we allow the values $1, 2, 3$ and 4. While, likewise the application of Section 3.3.3, the hyperparameters (α_s, β_s) on the precisions σ_j^{-2}'s, were set equal to $(2, 1)$, $(3, 2)$ and $(5, 4)$.

Looking at the histograms in Figures 3.6.1, we can see that for all combinations of α_c and (α_s, β_s), the number of discriminating variables is moving between 20 and 30. The posterior probabilities of inclusion for the 50 variables (Figure 3.6.2) show clearly, how the non-conjugate case succeeds in identifying the 20 discriminating variables. Finally, Figure 3.6.3 shows that the 4 clusters can be recovered regardless the choice of α_c and (α_s, β_s).

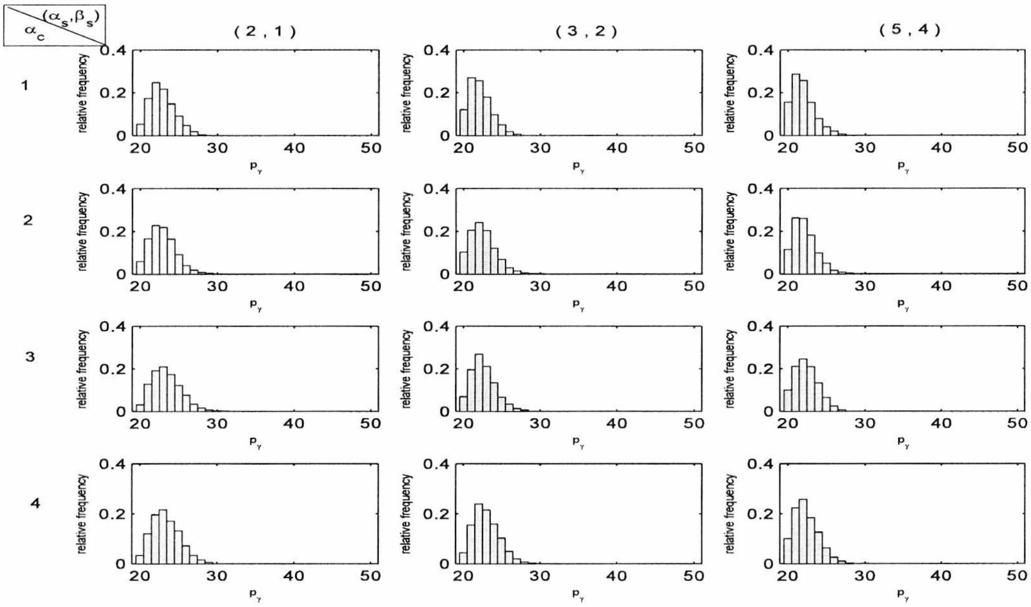

Figure 3.6.1: Non-Conjugate model for the simulated data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

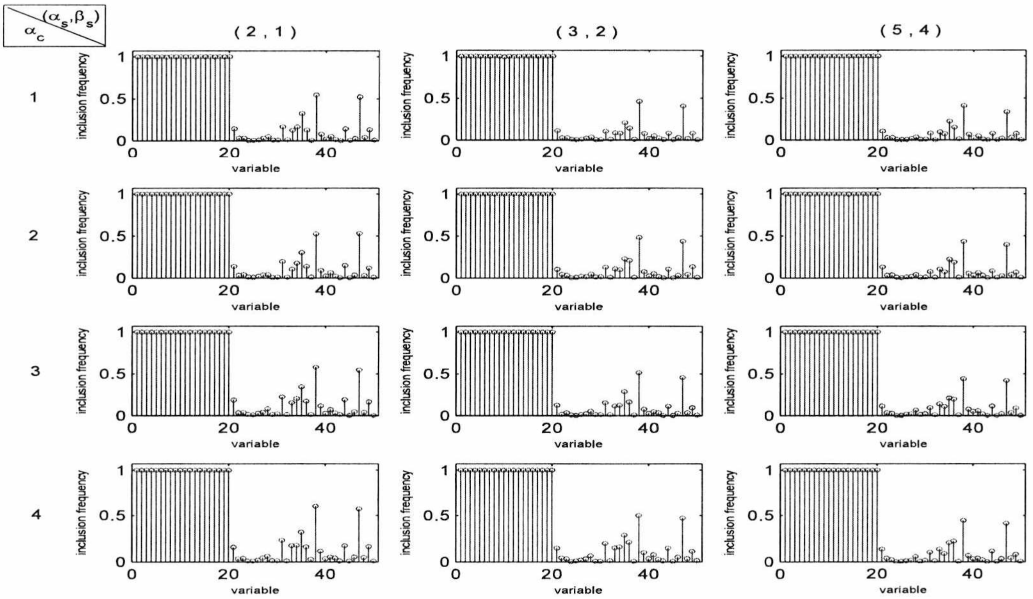

Figure 3.6.2: Non-Conjugate model for the simulated data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

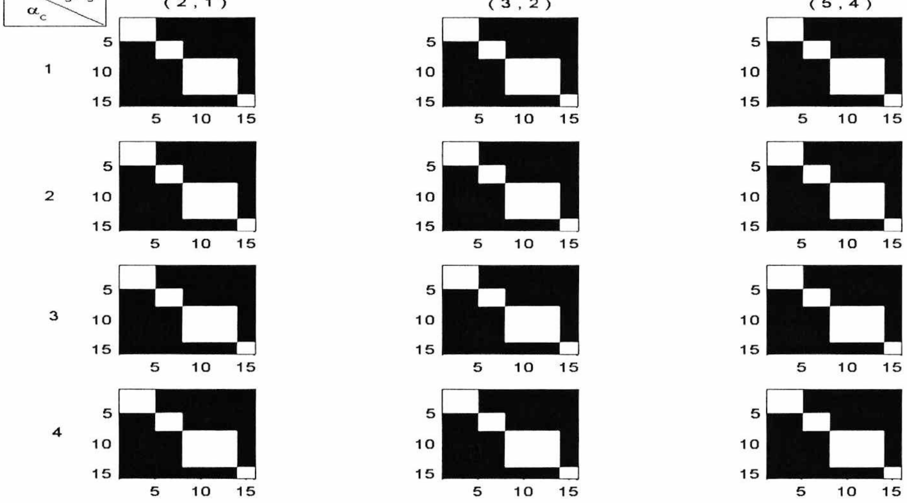

Figure 3.6.3: Non-Conjugate model for the simulated data set: Maps of the cluster allocations of the $n = 15$ observations for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

3.7 Conclusions

With regard to the model suggested in chapter 2 and the non-consistency of the results on account of the prior settings of its covariance structure, we introduced the idea of reconstructing the covariance structure. Starting with the Cluster Heterogeneity model, we considered the simplest case of setting the covariance matrices proportional to the Identity matrix. Application on our simulated data set showed a fair sensitivity on the choice of the hyperparameters under consideration; while hyperparameter h could still affect the recovery of the cluster structure of the data. Continuing, we built on the CH model and examined the Cluster-Variable Heterogeneity model, under which, the covariance matrices are set proportional to a diagonal matrix B. With a few hyperparameters to be controlled, we observe that h is again of major importance. Extending this we introduced a prior on the mean value m of the *Gamma* prior applied on the c_k's and eliminated the number of hyperparameters to be tuned. Results on different values of h though indicated the need of overcoming the strong impact of the hyperparameter. We, therefore, considered the model with an additional prior on h this time. Using a Markov chain we allowed sampling of h. Our application on the simulated data set encouraged the identification of the 20 discriminating variables and the recovery of the 4 clusters, regardless the prior settings. The latter can clearly state that we have managed to overcome the model's sensitivity on the hyperparameter h. Finally, we have developed the Non-Conjugate model. Using non-conjugate priors on the mean vectors and the covariance matrices, we managed to tackle any sensitivity previously imposed by the hyperparameters' settings, achieving satisfactory results on the variable selection and clustering tasks. We understand that both the CVH model with a prior on h as well as the Non-Conjugate model consist two robust models.

To conclude we give details on the computational time and acceptance rates of the models applied. Although, one would think that the Non-Conjugate model would impose an additional computational cost, it is the CVH model with prior on h and m (CVH_{hm}) model that does so. Nonetheless, the additional time is not dramatically larger. More specifically, on a two quad core server with 2.53Ghz CPUs, the CPU time for the CH, CVH, CVH_m and Non-Conjugate models was 3 hours, while for the application of the CVH_{hm} model we needed 3 hours and 30 minutes. Finally, in Table 3 we give the average acceptance rates for the split, merge, birth and death moves of the RJ sampler for all five models applied.

	CH	CVH		CVH_m		CVH_{hm}		NC
		$h = 100$	$h = 1000$	$h = 100$	$h = 1000$	$(3, 200)$	$(3, 400)$	
Split	0.2627	0.6875	0.8244	0.5239	0.858	0.3613	0.3913	0.111
Merge	0.2412	0.085	0.0605	0.1621	0.0403	0.2232	0.2319	0.2294
Birth	0.3647	0.5776	0.7064	0.5246	0.7259	0.4872	0.49	0.4363
Death	0.7946	0.7279	0.5563	0.7490	0.4843	0.6979	0.7447	0.7144

Table 3: Acceptance rates for the split, merge, birth and death moves of the RJ sampler on the application of the CH, CVH, CVH_m, CVH_{hm} and Non-Conjugate models on the simulated data set.

Chapter 4

Applications To Real Data Sets

4.1 Prologue

In this final chapter we examine and compare the performance of the models with the different covariance structures described in chapter 3 on three real data sets. We begin with two well-known and widely used data sets, the Iris data and the Crabs, and we conclude with a data set of the special feature of a vast number of variables with a substantially smaller sample size, the arthritis data set.

4.2 Iris Data

With measurements on the sepal length, sepal width, petal length and petal width (cm), where both sepals and petals are modified leaves that form the perianth of the flower, we have the 150 observations of the Iris data allocated into three groups, the *Iris setosa*, *Iris versicolor* and *Iris virginica* (Anderson, 1935). Each of the groups contains 50 observations, however, it is known that *Iris versicolor* and *Iris virginica* overlap, while the information provided is not enough for one to recover the three categories as these have been formed using additional covariates not included in the current data set (Anderson, 1936).

Indeed, this latter fact has characterised the results of the application of the models described in chapters 2 and 3. In the graphs that follow we will see the suggested grouping and the variables selected, first from the model of chapter 2 for the two different computational approaches of sections 2.4 and 2.7, followed by the results for the various covariance structures of chapter 3. In all cases, under the assumption of unequal covariance structures, we used a starting point of 3 groups, with the cluster allocations resembling the true cluster structure of the Iris data, while all four covariates are being included in the model. We ran the algorithms for 30000 iterations with a burn-in of 10000.

Starting from the application of the model under chapter 2, we need to give a few details concerning the values of the various hyperparameters. With $\delta = 3$ and $\alpha_u = 2$, we allowed a maximum number of 10 clusters ($G_{max} = 10$), using a prior Dirichlet with parameter $\alpha_w = 1$. For the variable selection task on the other hand, we chose a and b such that $a + b = 2$ with the number of selected variables being a priori equal to 2. Under these settings we then examined the sensitivity of the model on the choice of h and c - recall here that we let $h_1 = h_0$ and $c = d$ and therefore we refer to those hyperparameters under a common h and c respectively.

Although the values of h and c played a crucial part in the application of the two approaches on the simulated data set (sections 2.5 and 2.7), we see here that the different values for h and c ($h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$), alter neither the resulting clustering nor the choice of discriminating variables. However, we shall not consider this event as a success, as the results are rather disappointing. In Figures 4.2.1 and 4.2.4, where we can see the total number of discriminating variables, we observe that under all cases all four variables are chosen as discriminating and included in the model. Figures 4.2.2 and 4.2.5, indicating the posterior probabilities of inclusion for all variables of the iris data set, support such a choice with all four variables having high probability of inclusion. However, the clustering maps in Figures 4.2.3 and 4.2.6 show that an ambiguity characterises the clustering of the observations, as no concrete clusters can be identified.

More specifically, the algorithms fail to see clusters in our data, forming one single cluster in which all 150 observations are assigned to with moderate probability (gray colour). We should recall here that the clustering map illustrates the posterior probabilities of $y_i = y_j, i, j = 1, ..., n$. This time, though, we use a colouring scale such that white corresponds to probability 0 and black indicates probability 1.

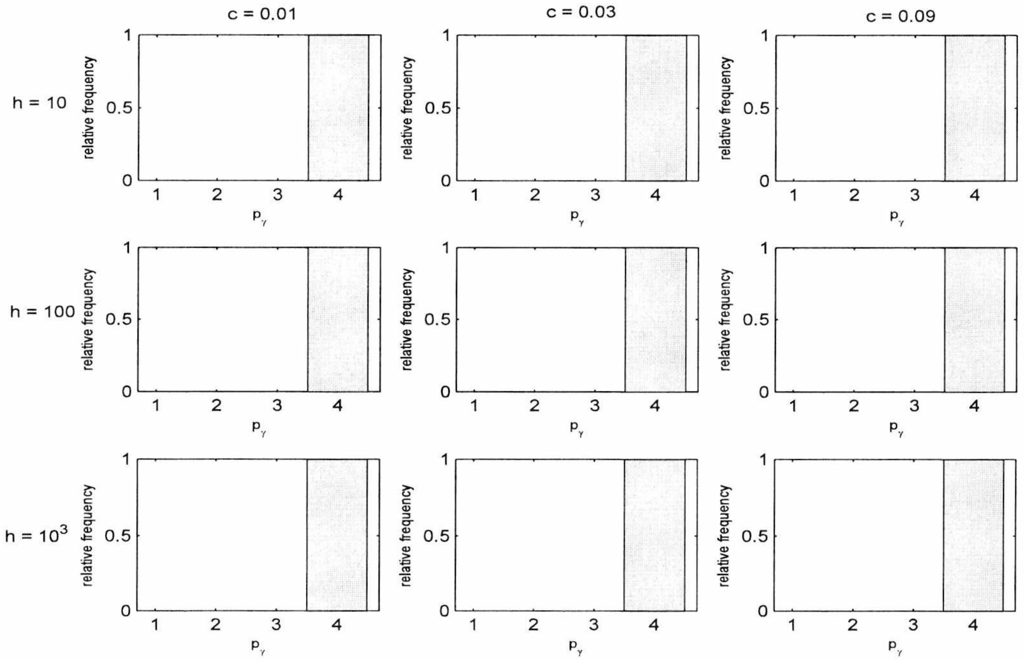

Figure 4.2.1: Iris data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 1$ and $\alpha_u = 6$, for the model of chapter 2.

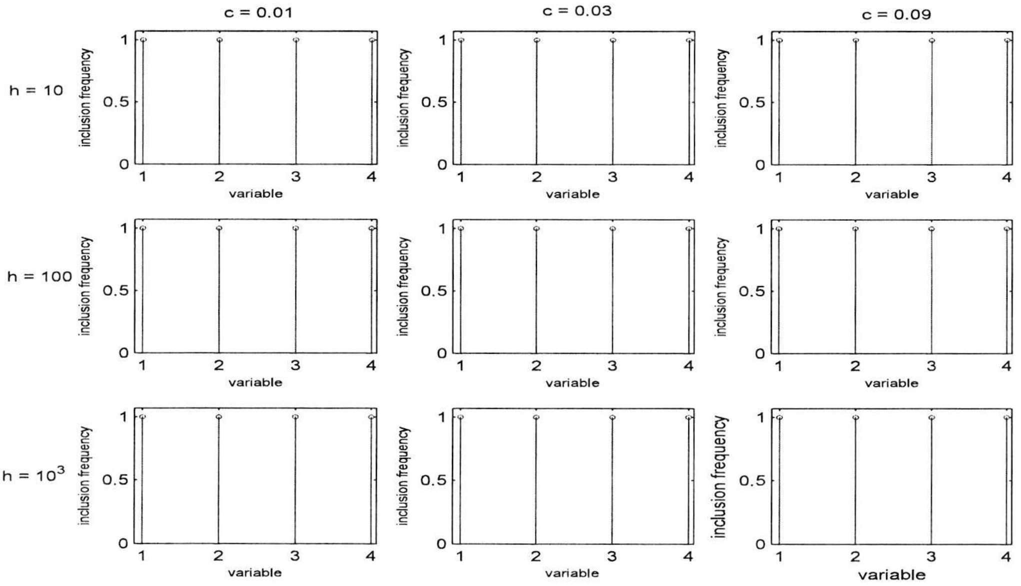

Figure 4.2.2: Iris data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 1$ and $\alpha_u = 6$, for the model of chapter 2.

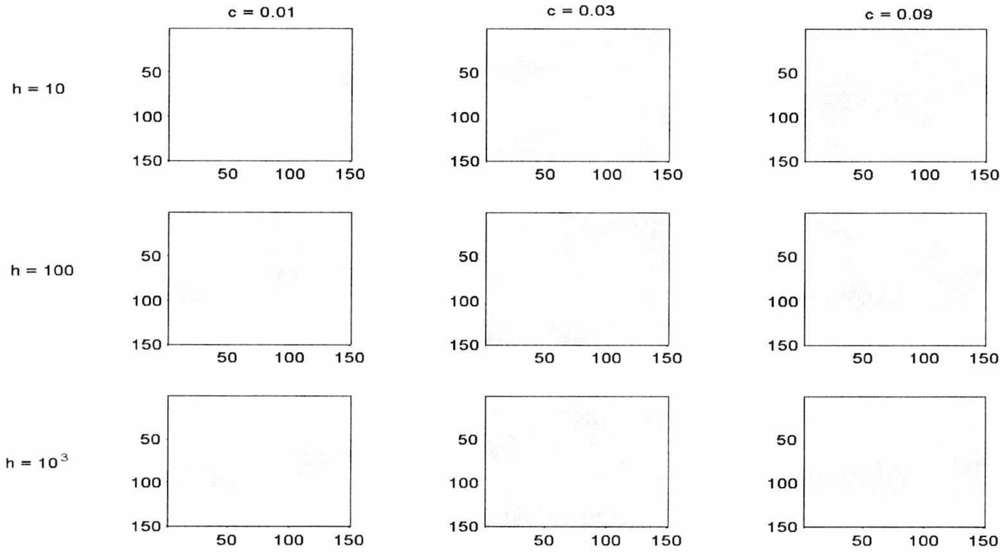

Figure 4.2.3: Iris data: Maps of the cluster allocations of the $n = 150$ observations for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 1$ and $\alpha_u = 6$, for the model of chapter 2.

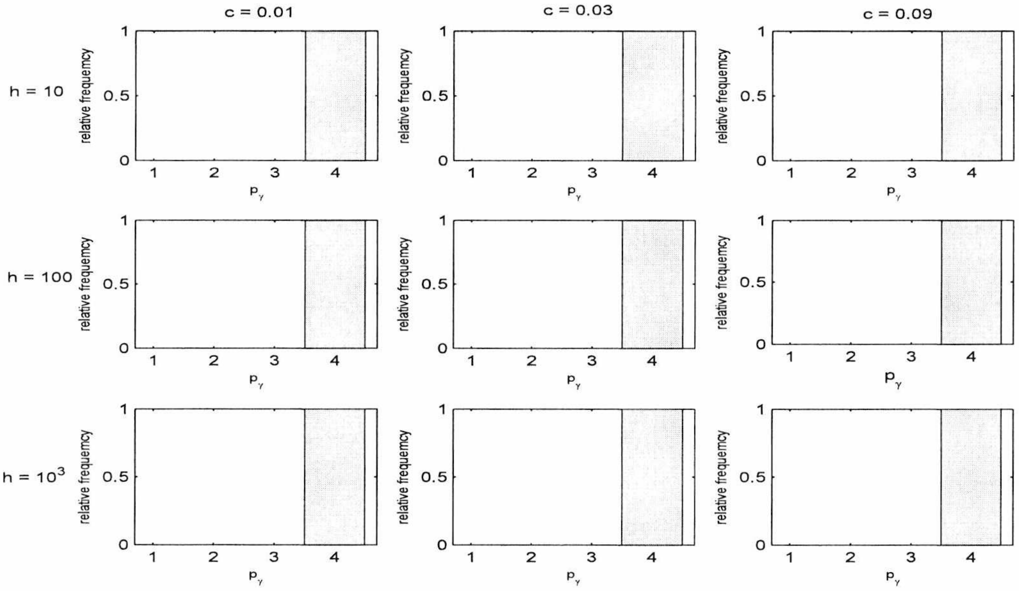

Figure 4.2.4: Iris data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

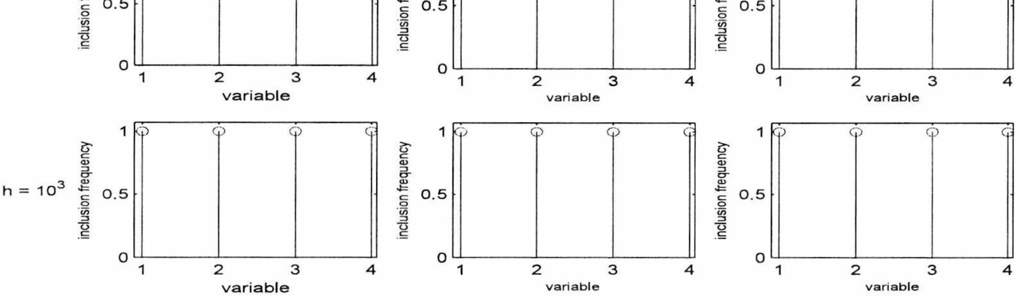

Figure 4.2.5: Iris data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

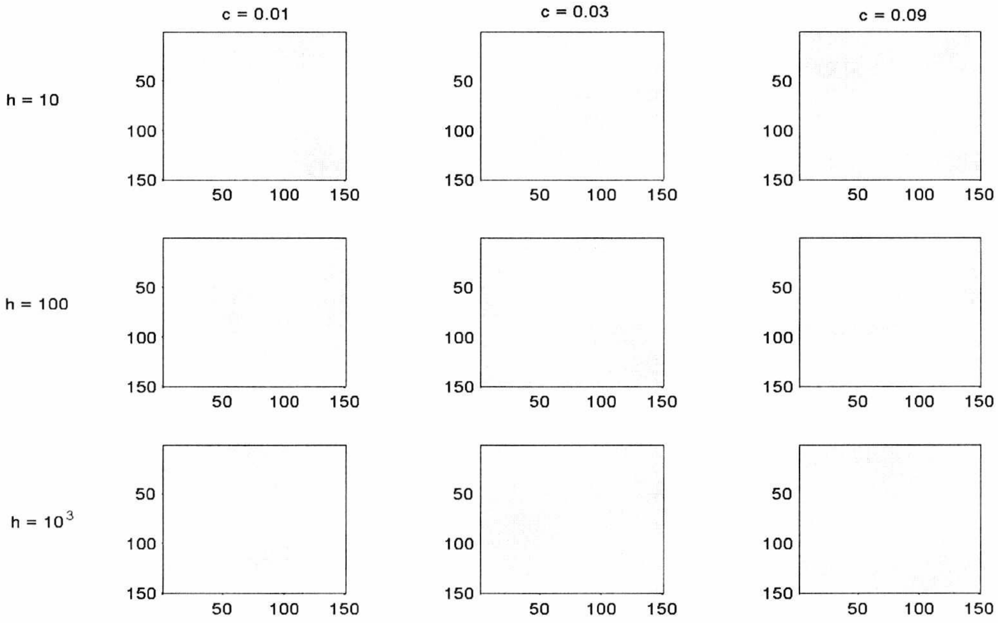

Figure 4.2.6: Iris data: Maps of the cluster allocations of the $n = 150$ observations for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

Moving to the Cluster Heterogeneity model of chapter 3 and using the same setting for $a, b, p_{prior}, G_{max}, \alpha_u$ and α_w, we examined the behaviour of the model for the values of $h = 100$ and $h = 1000$. We considered the hyperparameters of the Inverse Gamma distributions on the scalars c and d all equal and tried the values $1, 2$ and 1000.

Once again, looking at the total number of selected variables and their posterior probabilities (Figure 4.2.7 and 4.2.8), we see that all four variables are considered as discriminating, with the values of $h, \alpha_c, \beta_c, \alpha_d$ and β_d not affecting the selection of variables. However, in the clustering maps of Figure 4.2.9, the effect of $\alpha_c, \beta_c, \alpha_d$ and β_d is clearly illustrated. While values of $\alpha_c, \beta_c, \alpha_d$ and β_d being either 2 or 1000 comply with the consideration of a single cluster, we see that for $\alpha_c = \beta_c = \alpha_d = \beta_d = 1$ the algorithm picks up some sort of clustering. Clearly, for both cases of $h = 100$ and $h = 1000$ the resulting clusters are similar,

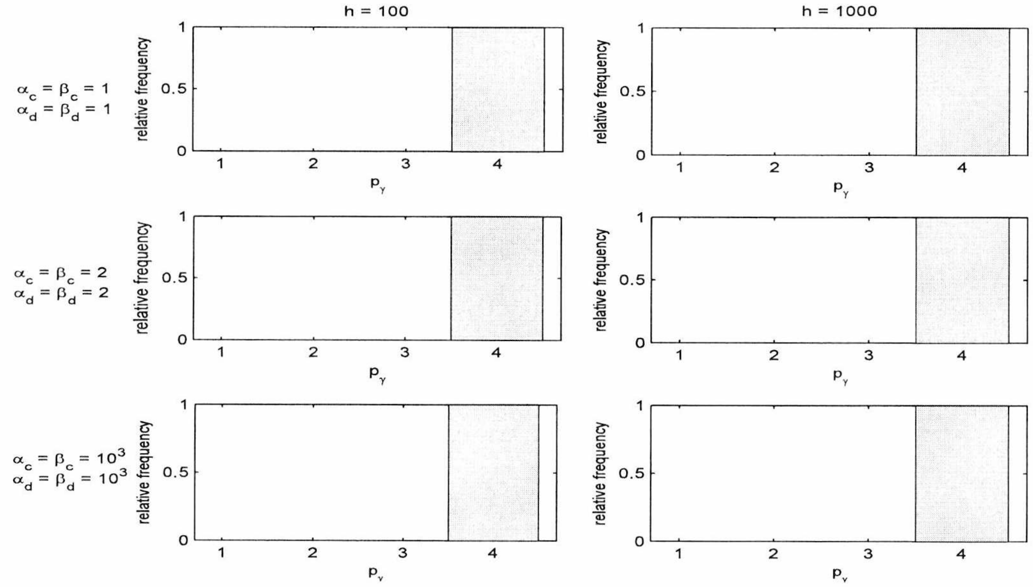

Figure 4.2.7: Cluster Heterogeneity for the Iris data set: Histograms of the total number of discriminating variables, p_γ, assuming unequal covariance matrices, for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ with values $1, 2$, and 1000.

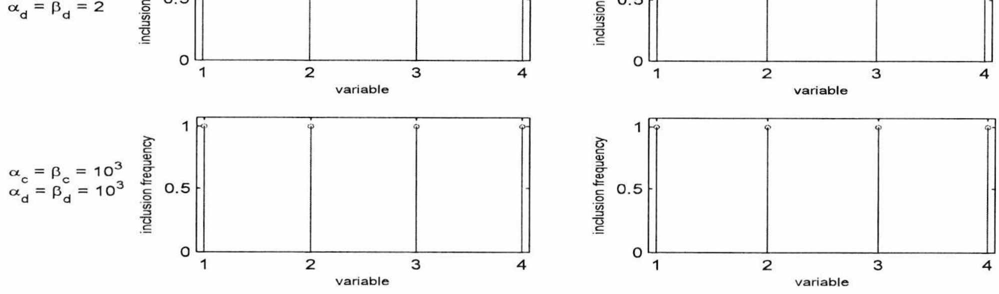

Figure 4.2.8: Cluster Heterogeneity for the Iris data set: Marginal Posterior Probabilities of the variables included in the model, assuming unequal covariance matrices, for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ with values $1, 2$, and 1000.

with the *Iris setosa* category being fully recovered. While the first 50 observations are correctly assigned to the same cluster, we can also see that the last 50 observations, originally allocated to *Iris virginica*, successfully form a separate cluster as well, with only a few observations being missgrouped and actually allocated together with the observations of the first cluster. However, the cluster allocations of the remaining observations have an interesting turnout; part of them are assigned to the *Iris setosa* group, while the rest comply with the observations of *Iris virginica*.

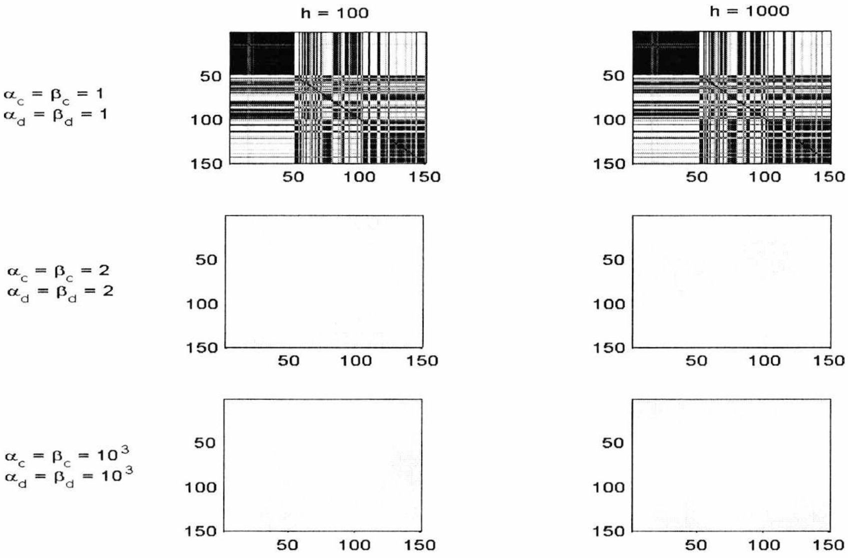

Figure 4.2.9: Cluster Heterogeneity for the Iris data set: Maps of the cluster allocations of the $n = 150$ observations for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ equal to $1, 2,$ and 1000, under the assumption of unequal covariance matrices.

In other words, we can tell that under the assumption of a simple covariance structure, CH model, the algorithm identifies two clusters. That is, the *Iris setosa* and *Iris virginica*, with the observations originally allocated in *Iris versicolor* being allocated to either of the two.

Coming to the Cluster - Variable Heterogeneity model, using a hyperprior on m, we try the values $\alpha_c = (2, 3, 4)$ and the pairs $(\alpha_s, \beta_s) = [(2, 1), (3, 2), (5, 4)]$, for both the cases of $h = 100$ and $h = 1000$. The number of discriminating variables, together with their posterior probabilities of inclusion and the suggested cluster structure of our data for the case of $h = 100$ in Figures 4.2.10, 4.2.11 and 4.2.12, indicate that under all cases we take similar results for both the variable selection and clustering tasks. Interestingly, a value of $h = 1000$, produces similar results. We give the corresponding Figures (Figures C.0.13 - C.0.13) in Appendix C.

With all variables being selected to be included in the model with high probability (Figure 4.2.11), we see that the two clusters *Iris versicolor* and *Iris virginica* overlap as expected, but this time, unlike the case of the CH model, they are well separated from the first group, which, on the other hand, is not fully recovered. With the colouring map indicating a split of the *Iris setosa* into further smaller groups, taking a closer look, we could tell that if we reorder the first 50 observations we could clearly see the formulation of only two smaller groups.

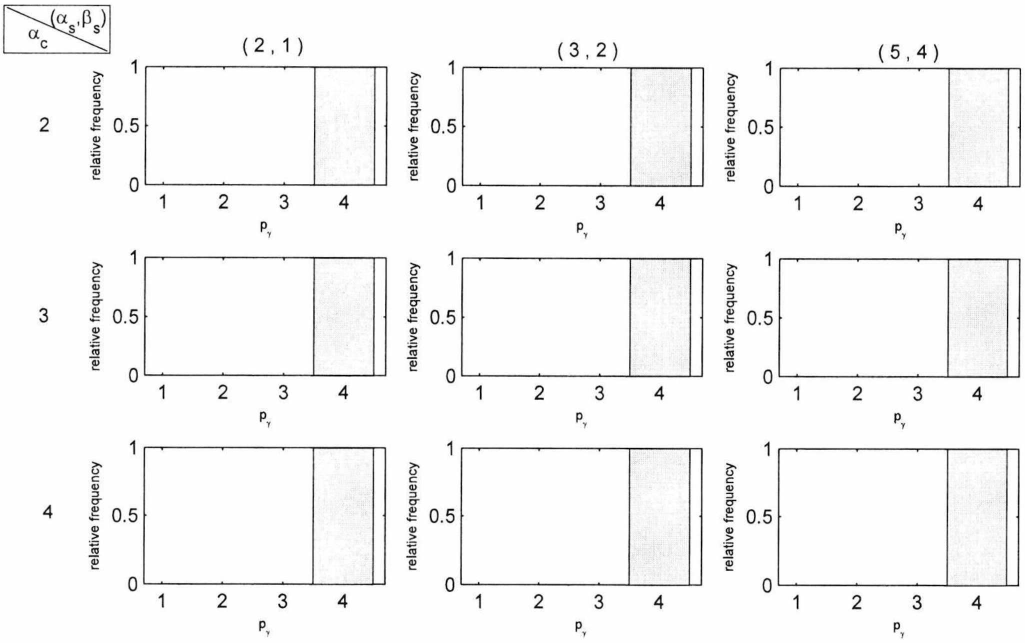

Figure 4.2.10: Cluster-Variable Heterogeneity with a prior on m for the Iris data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 100$.

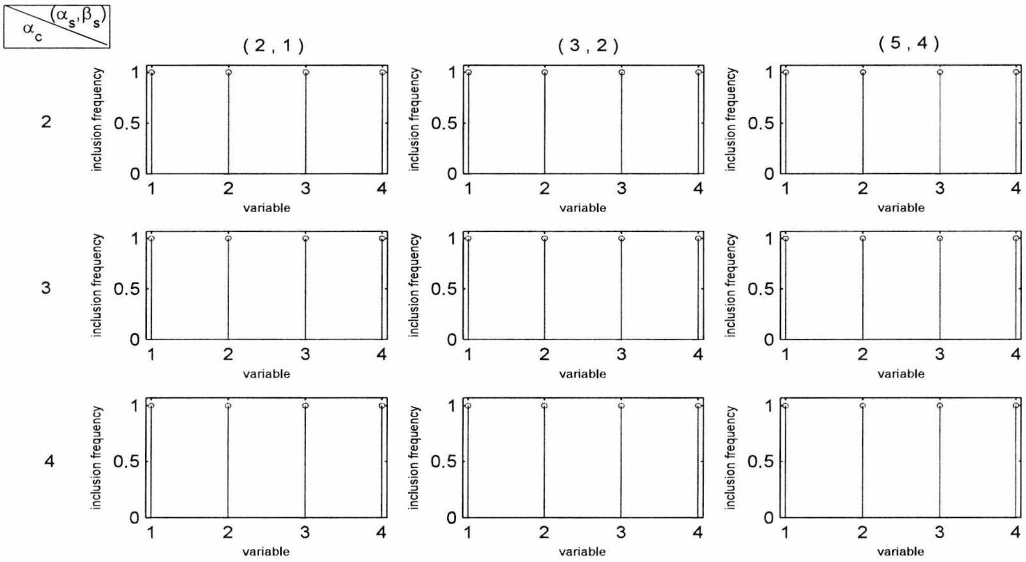

Figure 4.2.11: Cluster-Variable Heterogeneity with a prior on m for the Iris data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 100$.

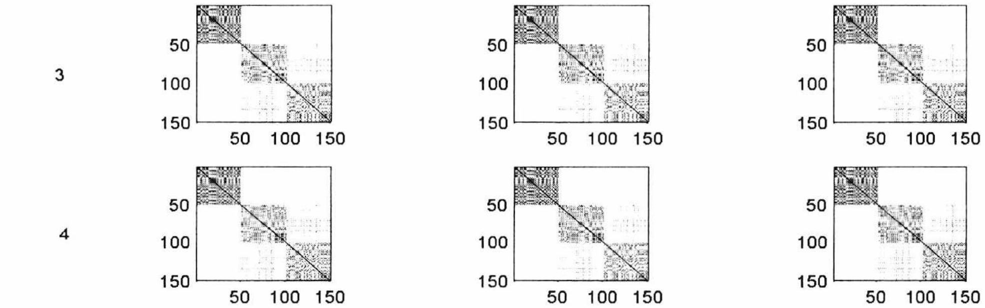

Figure 4.2.12: Cluster-Variable Heterogeneity with a prior on m for the Iris data set: Maps of the cluster allocations of the $n = 150$ observations for various combinations of (α_s, β_s) and α_c assuming heterogeneity and setting $h = 100$.

Although the model does not manage to identify the *Iris setosa* species as a single cluster, but rather splits it into two smaller ones, the robustness of the model is rather encouraging, indicating that there is no need for additional care on the choice of any of the hyperparameters, not even the once crucial h.

Even though h does not seem to have an impact on the results so far, it would be interesting to examine the case of setting a hyperprior on h, allowing inference on h. Trying the same combinations of α_c and (α_s, β_s), we used an Inverse Gamma prior on h with parameters $(3, 200)$ and $(3, 400)$. Besides the number of discriminating variables with the probabilities of inclusion for the four variables, as well as the cluster allocations, we estimated the value of h. With parameters $(3, 200)$ and $(3, 400)$ giving similar results, we provide Figures for the case of an Inverse Gamma with $(3, 200)$ (Figures 4.2.13 - 4.2.16). Figures for the case of $h \sim IG\,(3, 400)$ are given in Appendix C (Figures C.0.16 - C.0.19).

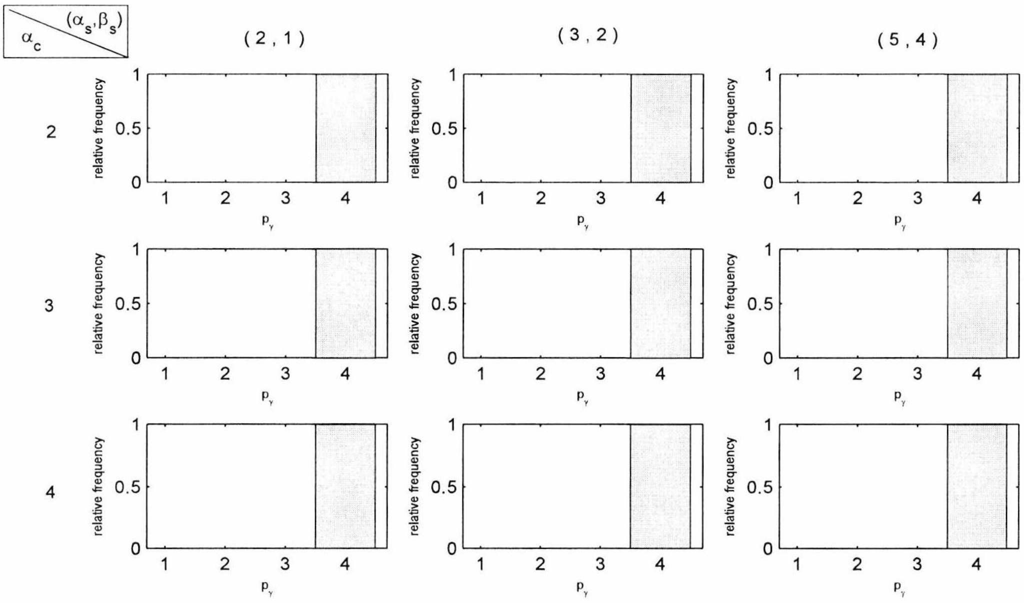

Figure 4.2.13: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

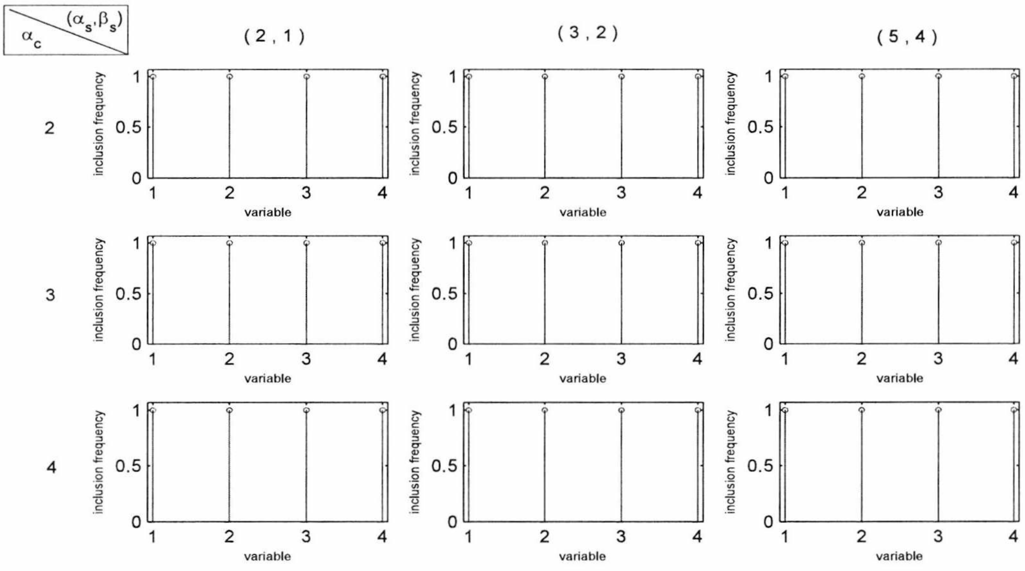

Figure 4.2.14: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

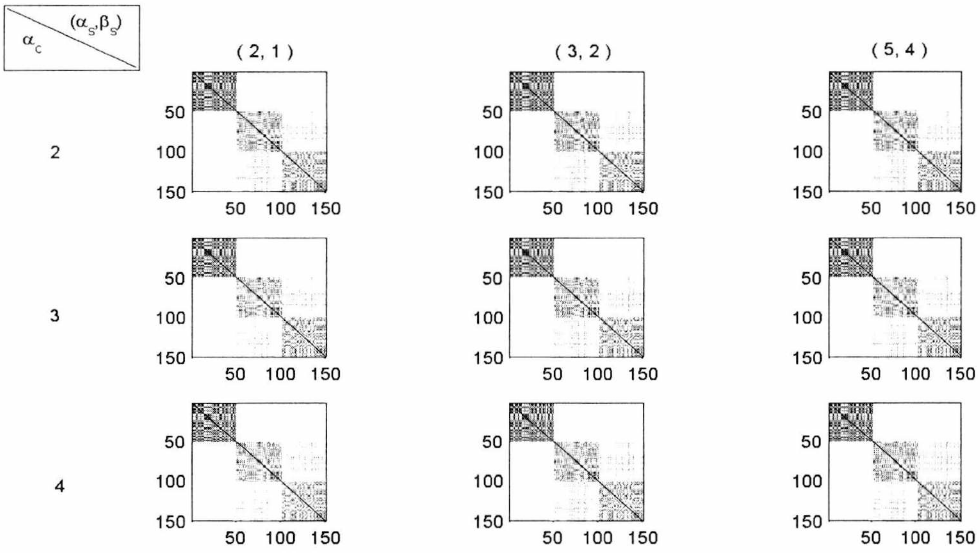

Figure 4.2.15: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Maps of the cluster allocations of the $n = 150$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

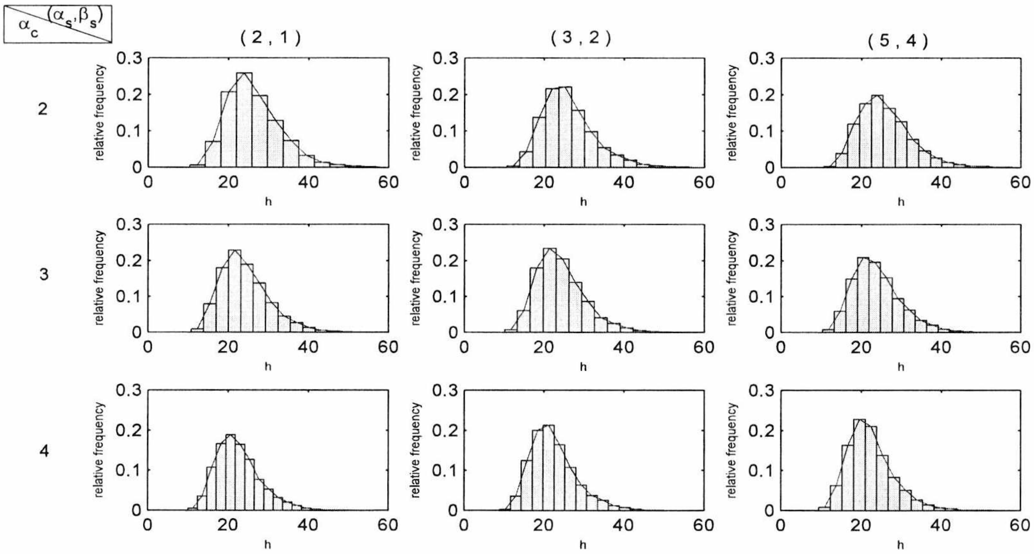

Figure 4.2.16: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

It is very interesting to observe that, for all cases we experimented on, the estimated value for h is roughly 20, regardless the settings of its hyperprior. But most importantly, together with the full set of variables being identified as discriminating, $h = 20$ gives cluster allocations that resemble the cluster matrices for the CVH model with a prior on m (Figure 4.2.12), suggesting the split of the first group (*Iris setosa*) into two smaller ones and finding that *Iris versicolor* and *Iris virginica* overlap in the exact same way as under the consideration of a preset value of h to 100 and 1000. That is a further indication that for the Iris data the hyperparameter h is not of great importance.

Finally, we examine the case of non-conjugacy. Setting $\alpha_c = 2, 3$ and 4, combined with the values $(\alpha_s, \beta_s) = (2, 1), (3, 2)$ and $(5, 4)$, we observe that for all cases, except for $\alpha_c = 4$ with $(\alpha_s, \beta_s) = (2, 1)$ and $(3, 2)$, all variables are included in the model (Figure 4.2.17) with high probabilities of inclusion (see Figure 4.2.18).

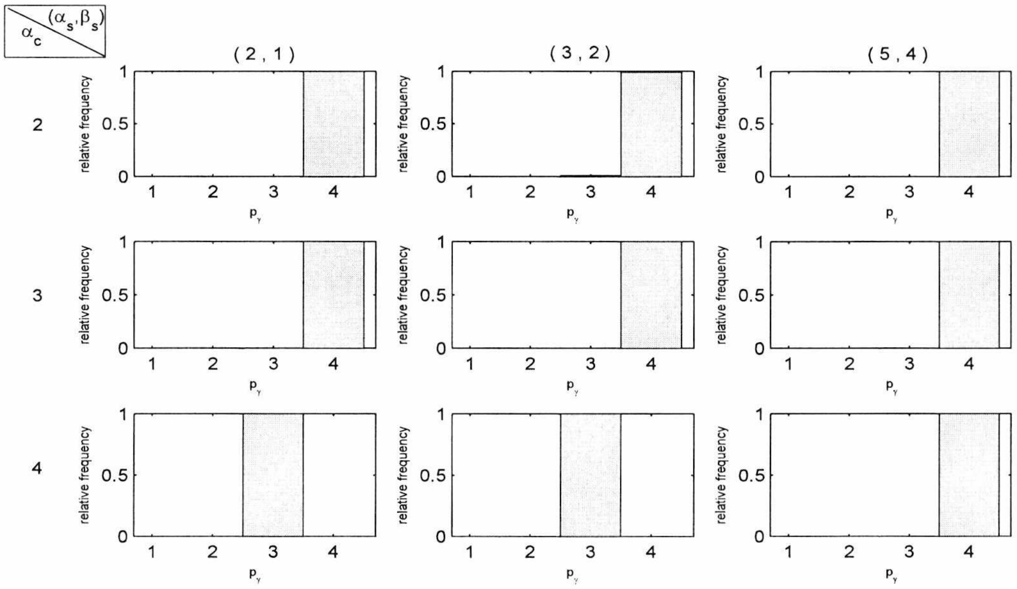

Figure 4.2.17: Non-Conjugate model for the Iris data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

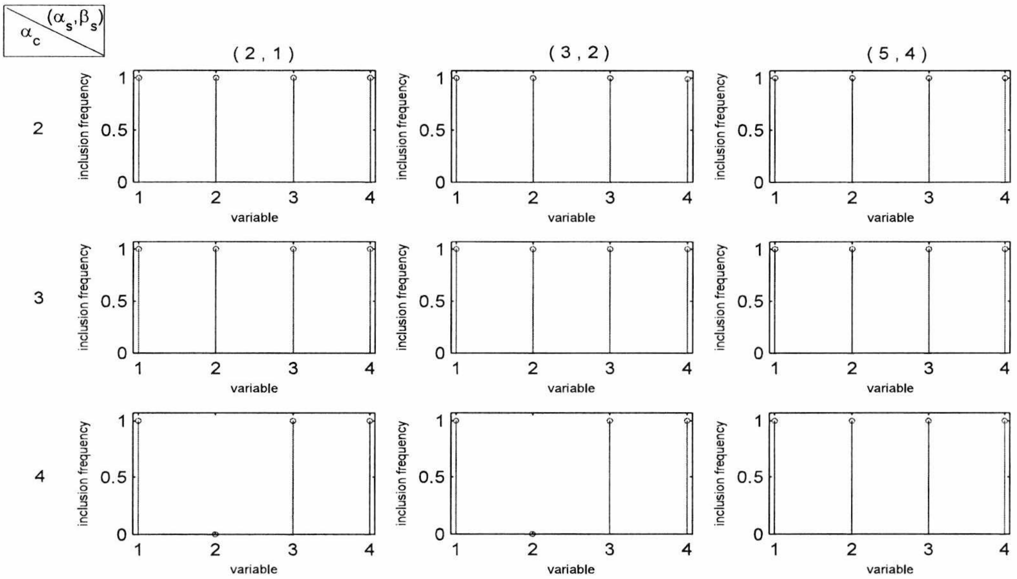

Figure 4.2.18: Non-Conjugate model for the Iris data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

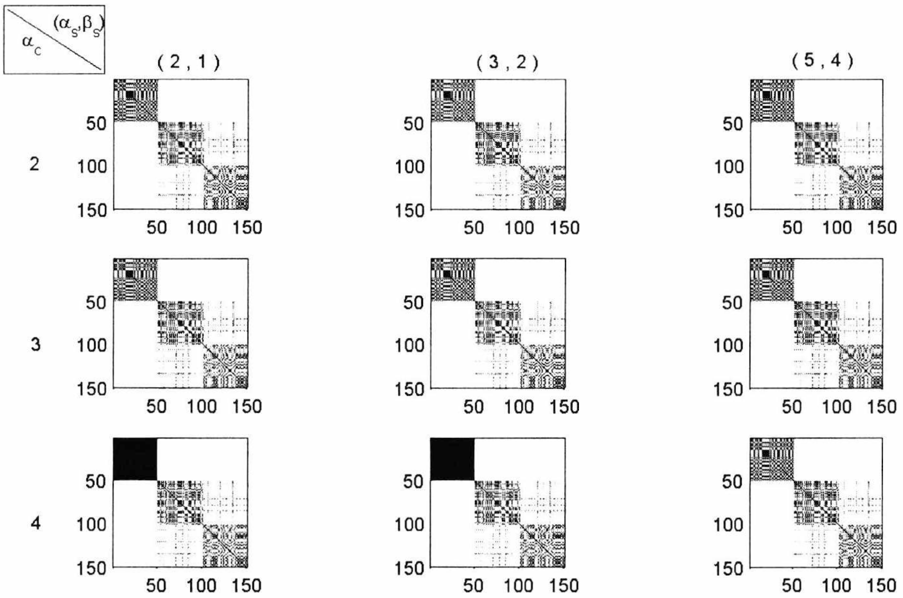

Figure 4.2.19: Non-Conjugate model for the Iris data set: Maps of the cluster allocations of the $n = 150$ observations for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

The corresponding suggested clusters (Figure 4.2.19) resemble the clustering structure proposed by the methods examined so far, splitting the *Iris setosa* into two smaller groups, with the *Iris versicolor* and *Iris virginica* groups overlapping. However, looking at the two exceptional cases of $\alpha_c = 4$ with $(\alpha_s, \beta_s) = (2,1)$ and $(3,2)$, while 3 out of 4 variables are selected, with the posterior probabilities of Figure 4.2.18 indicating the second variable as non-discriminating, we observe that the *Iris setosa* category is fully recovered.

On summarising, for all models applied on the Iris data set we had difficulties in identifying the 3 species of flowers. More specifically, the problem is focused on the observations originally allocated to the Iris versicolor and the Iris virginica groups. Under all cases observations from the versicolor group are being allocated to the virginica one and vice versa. An explanation of such misclassification could be a possible indication of overlapping groups. Indeed, performing a principal components analysis and plotting the first two principal components (Figure

4.2.20) one can observe such a phenomenon. The Iris setosa group is well separated from the other observations, however, observations from versicolor and virginica are very close together, indicating that the two groups clearly overlap. We therefore understand that it is not unlikely that cluster analysis can have problems in distinguishing the two groups, a remark present throughout our applications.

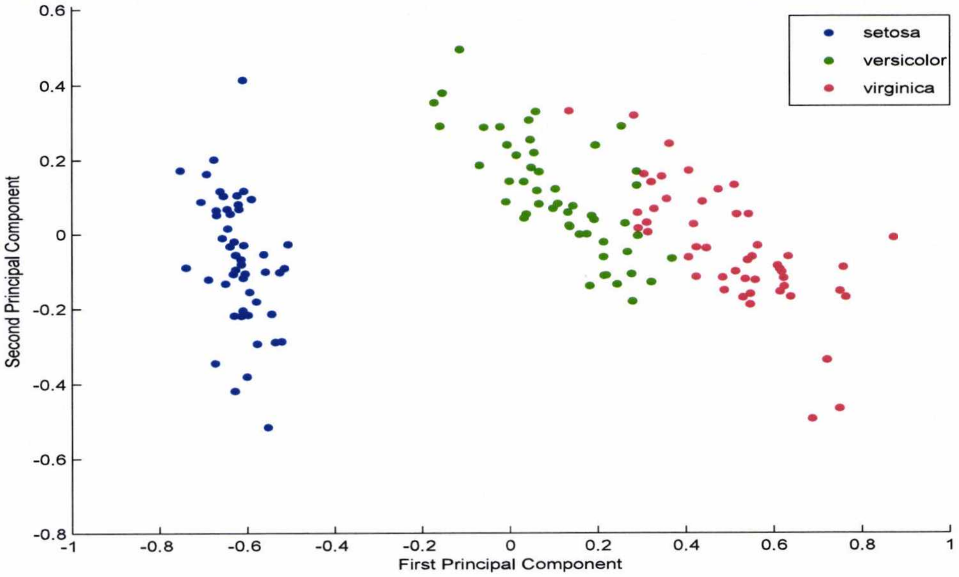

Figure 4.2.20: Plot of the two first principal components of the Iris data.

4.3 Crabs

The Crabs data set consists of 5 variables measured on 200 observations that form 4 groups. The variables refer to the width of frontal lip, rear width, length along the midline of the carapace, maximum width of the carapace and body length (all in mm), while there are two species of crabs, the orange and the blue. Considering the female and male subcategories, we finally have four groups, the orange female, orange male, blue female and blue male, with 50 observations in each of them.

However, the five variables of the crabs data set are highly correlated (see

Figure 4.3.1) and thus working in the principal components' space instead has been commonly preferred [Yau and Holmes (2011), Raftery and Dean (2006)]. We therefore performed a principal components analysis on the standardised data and applied the variables selection and clustering procedure on the principal components' space, allowing the algorithm to pick the most important principal components. We chose 10 as the maximum number of clusters and initiated the algorithm considering all principal components included in the model, when the original formulation of groups has been used as the starting point for the cluster allocation vector. We finally allowed $\alpha_u = 2$ and $\alpha_w = 1$ and run 30000 iterations with 10000 as the burn in period, assuming unequal covariance matrices along the groups.

Looking at the pairs plot of the five principal components in Figure 4.3.2, we can see the four groups being clearly formed on the space of the second and third components. However, in the figures that follow we will see that for most of the cases we tried, all principal components are selected from the various models applied, with only a few cases selecting a smaller number of components, which on the other hand do not provide a satisfactory clustering. But let's examine each case separately.

Also, looking briefly at the results from the application of the two computational approaches of the model in chapter 2 and the histograms for the total number of discriminating variables included in the model (Figures 4.3.3, 4.3.6), we see that for both approaches the algorithm oscillates between 3 and 5 principal components. The clustering maps for the two approaches in Figures 4.3.5 and 4.3.8 respectively, under most cases, suggest a single cluster. Despite the two exceptions of $c = 0.09$ with $h = 10$ and $h = 100$, when using the SAMS sampler (section 2.7), the four groups of the crabs data cannot be identified.

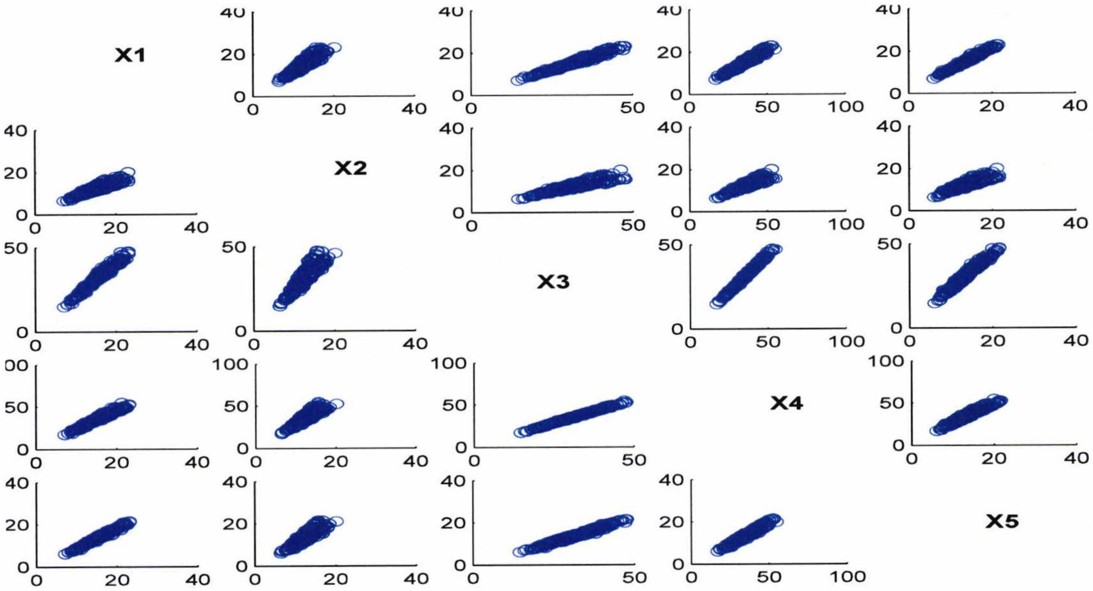

Figure 4.3.1: Pairs Plot of the five variables of the Crabs data set.

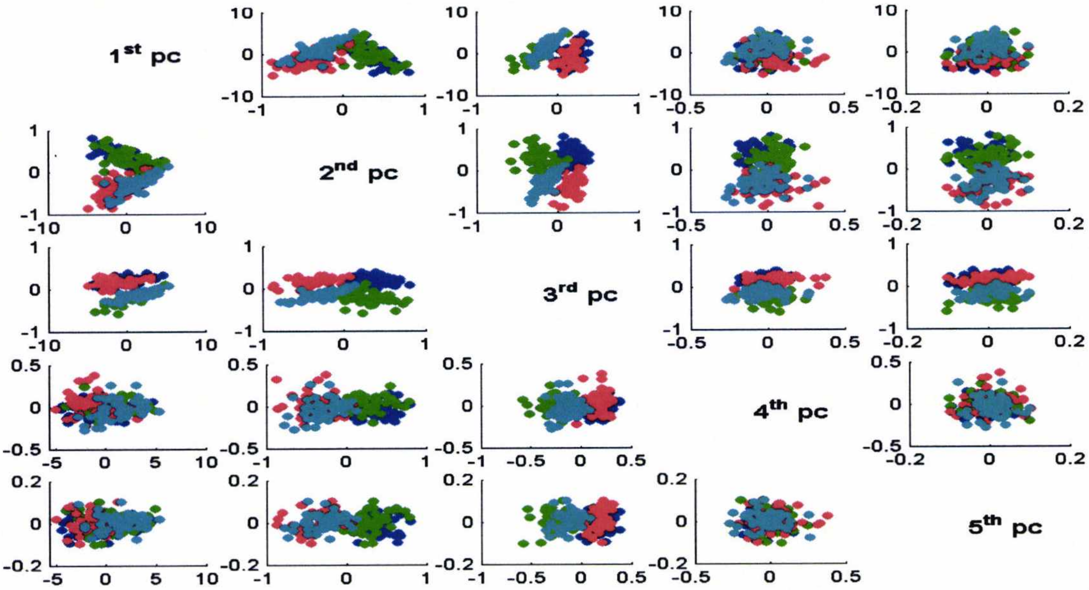

Figure 4.3.2: Pairs Plot for the 5 principal components of the standardised Crabs data set.

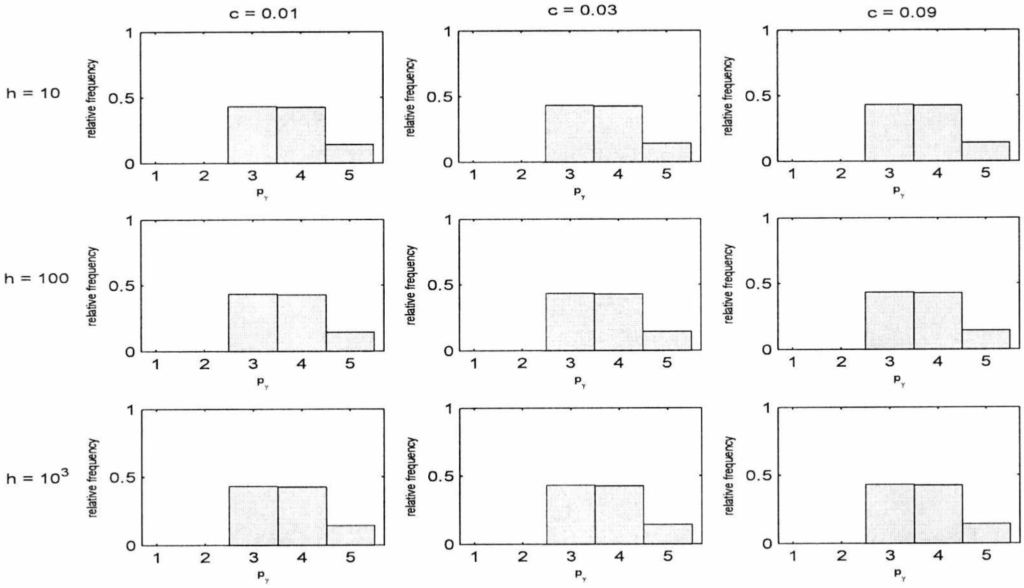

Figure 4.3.3: Principal components of Crabs: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices.

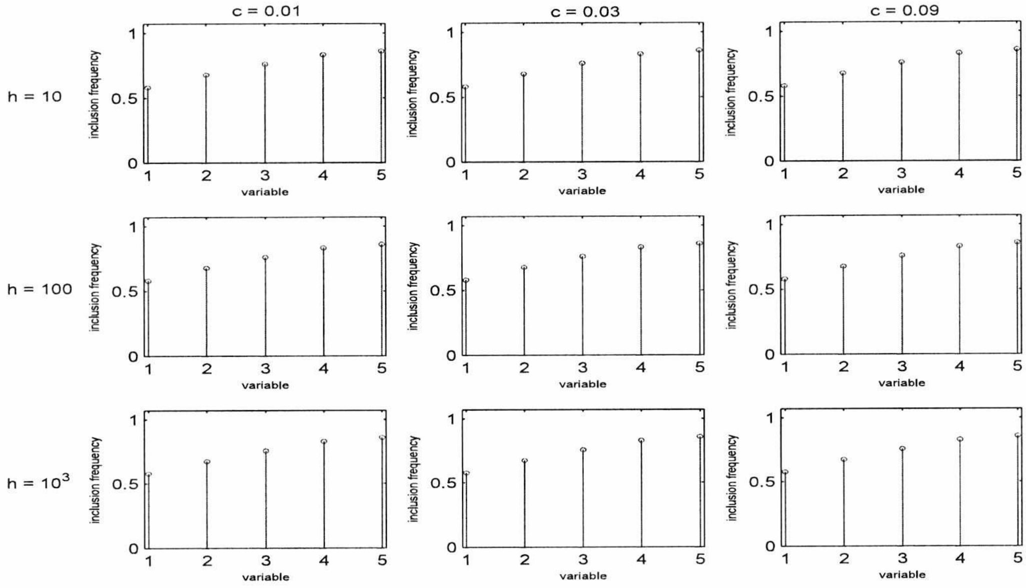

Figure 4.3.4: Principal components of Crabs: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices.

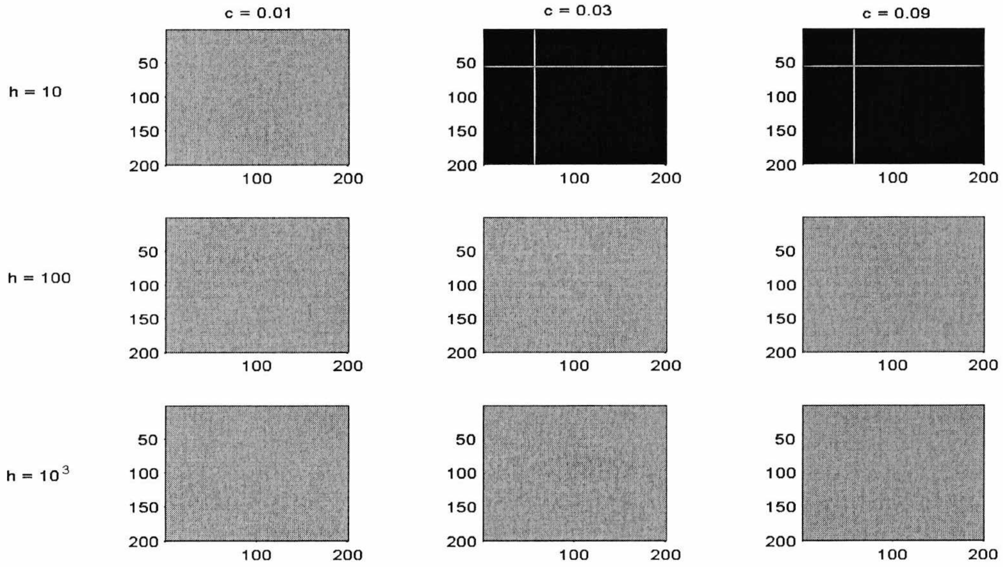

Figure 4.3.5: Principal components of Crabs: Maps of the cluster allocations of the $n = 200$ observations for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices.

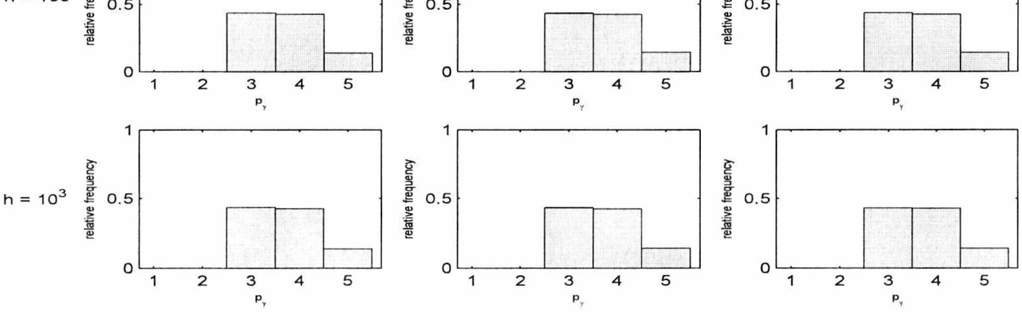

Figure 4.3.6: Principal components of Crabs: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

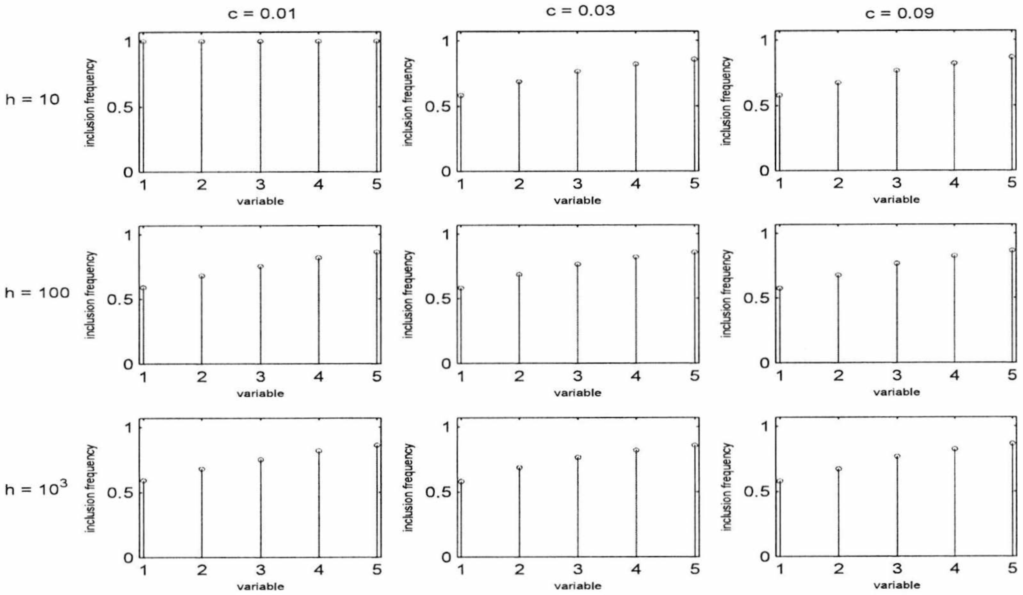

Figure 4.3.7: Principal components of Crabs: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

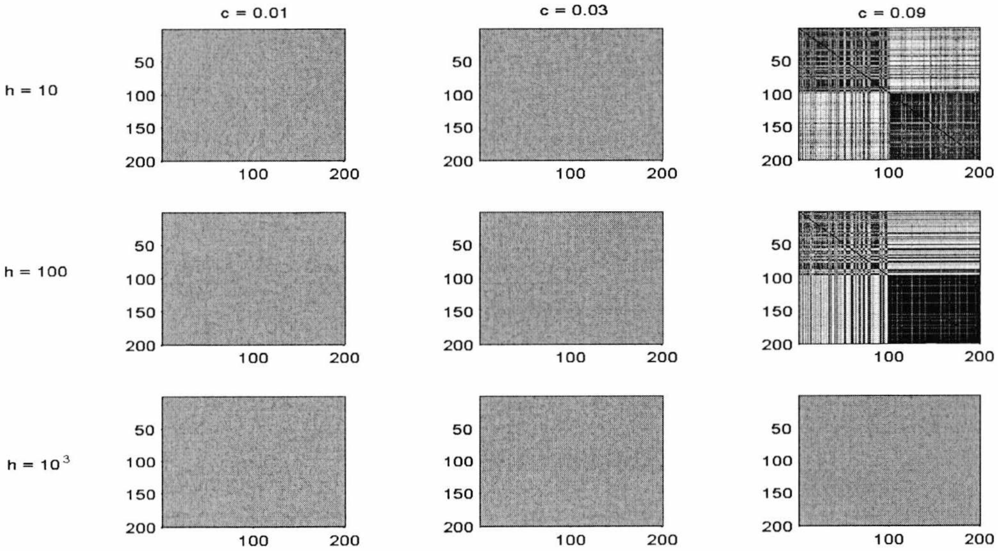

Figure 4.3.8: Principal components of Crabs: Maps of the cluster allocations of the $n = 200$ observations for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and using split/merge move of the SAMS sampler.

In the case of considering the CH model, the number of selected principal components in Figure 4.3.9 varies. According to the posterior probabilities of inclusion in Figure 4.3.10 though, we observe that the first principal is always included in the model. The clustering maps of Figure 4.3.11 on the other hand, do not indicate any particular cluster structure of the data. Only in the case of selecting 4 principal components, and that is for $h = 100$ and $\alpha_c, \beta_c, \alpha_d, \beta_d = 2$, the algorithm identifies two clusters, which with a few exceptions, e.g. the $50^{th}, 101^{st}, 150^{th}$ etc observations, we could say that these are the wider groups of orange and blue crabs. In other words, we could tell that for this specific case, although the model manages to identify the categories of sex, it fails in distinguishing the colour groups.

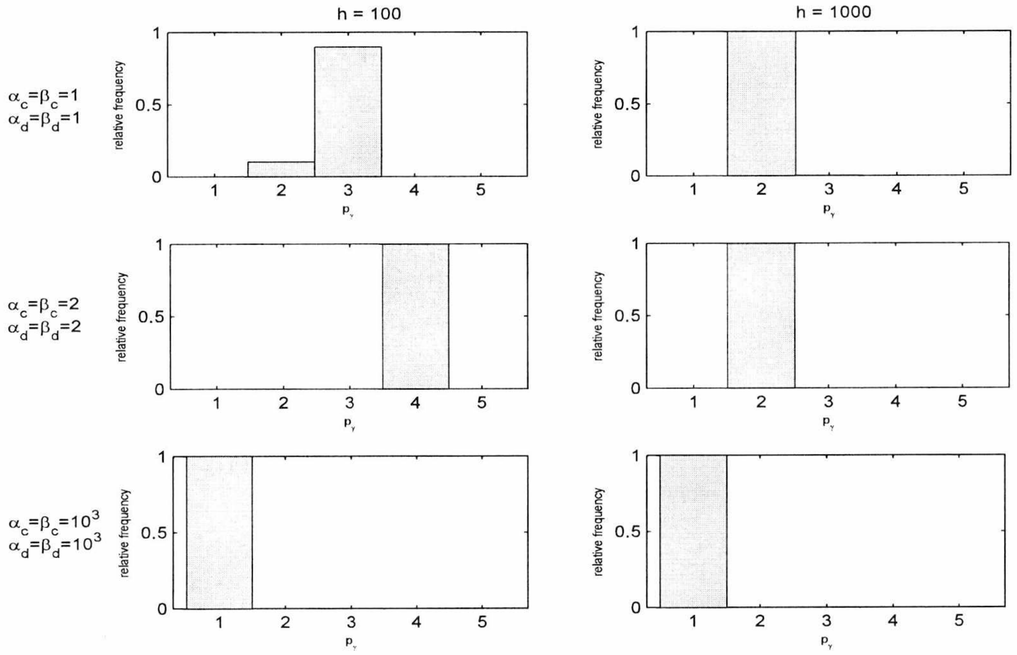

Figure 4.3.9: Cluster Heterogeneity for the principal components of the Crabs data set: Histograms of the total number of discriminating variables, p_γ, assuming unequal covariance matrices, for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ with values $1, 2$, and 1000.

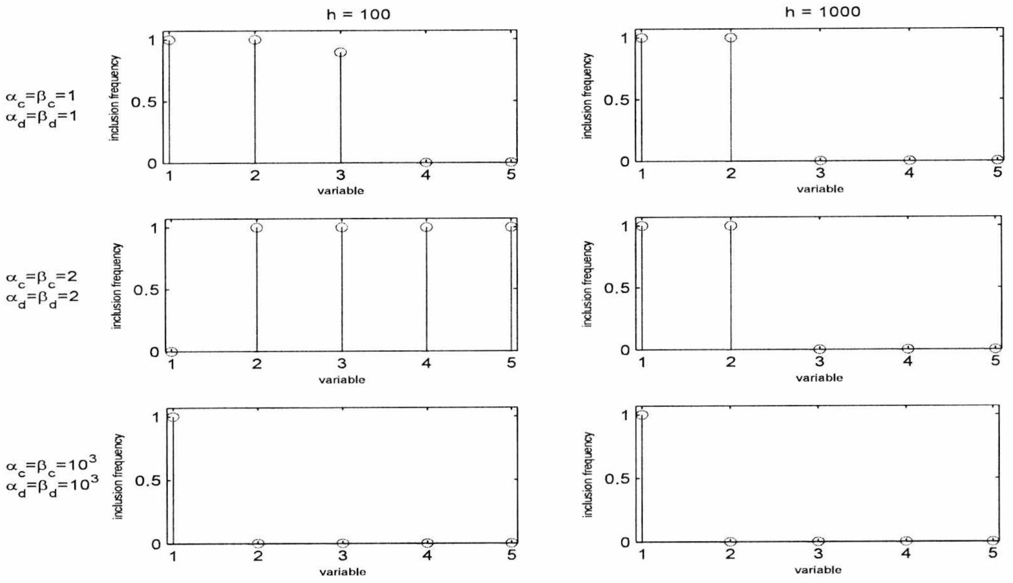

Figure 4.3.10: Cluster Heterogeneity for the principal components of the Crabs data set: Marginal Posterior Probabilities of the variables included in the model, assuming unequal covariance matrices, for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ with values $1, 2$, and 1000.

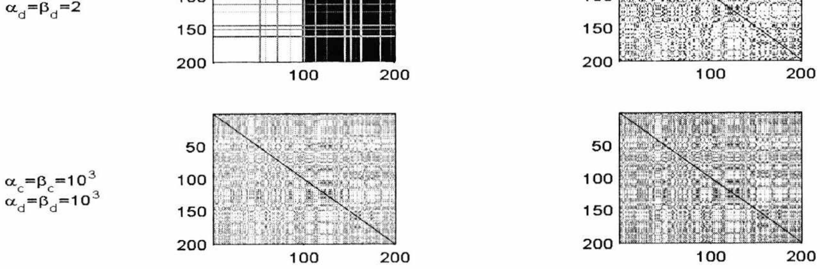

Figure 4.3.11: Cluster Heterogeneity for the Principal Components of the Crabs data set: Maps of the cluster allocations of the $n = 200$ observations for $h = 10, 100, 1000$ and $\alpha_c = \beta_c = \alpha_d = \beta_d$ equal to $1, 2$, and 1000, under the assumption of unequal covariance matrices.

Further, applying the CVH model having a prior on m, we tried different combinations for the shape parameter α_c and the pair (α_s, β_s), trying two values for h: 100 and a 1000. In Figure 4.3.12 and for $h = 100$, we observe that the number of selected principal components once again varies. Assuming the case of $h = 1000$ on the contrary (Figure 4.3.13), most combinations suggest the inclusion of all five principal components in the model. The suggested groupings though, show a lot of similarities regardless the choice of h or α_c, (α_s, β_s), with an interesting feature capturing ones attention (Figures 4.3.16, 4.3.17). We observe that although the pattern of four groups can be identified, the few cases of misgrouped observations correspond to the merging of observations from the same colour group but different sex. In particular, observations of the second group (orange male), tend to be merged with the observations of the fourth group (orange female).

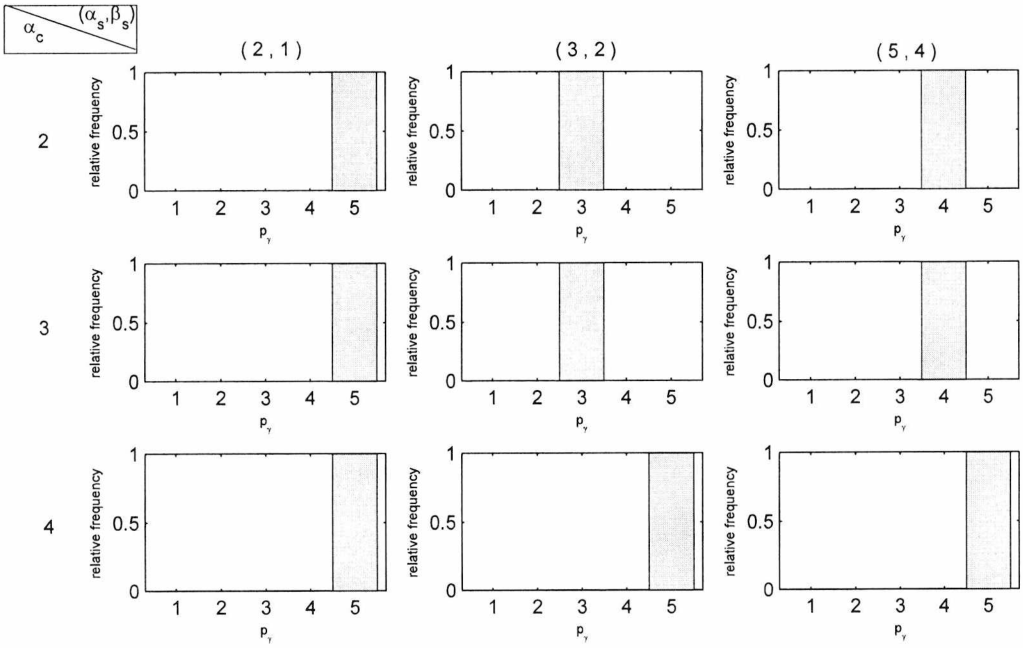

Figure 4.3.12: Cluster-Variable Heterogeneity with a prior on m for the principal components of the Crabs data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 100$.

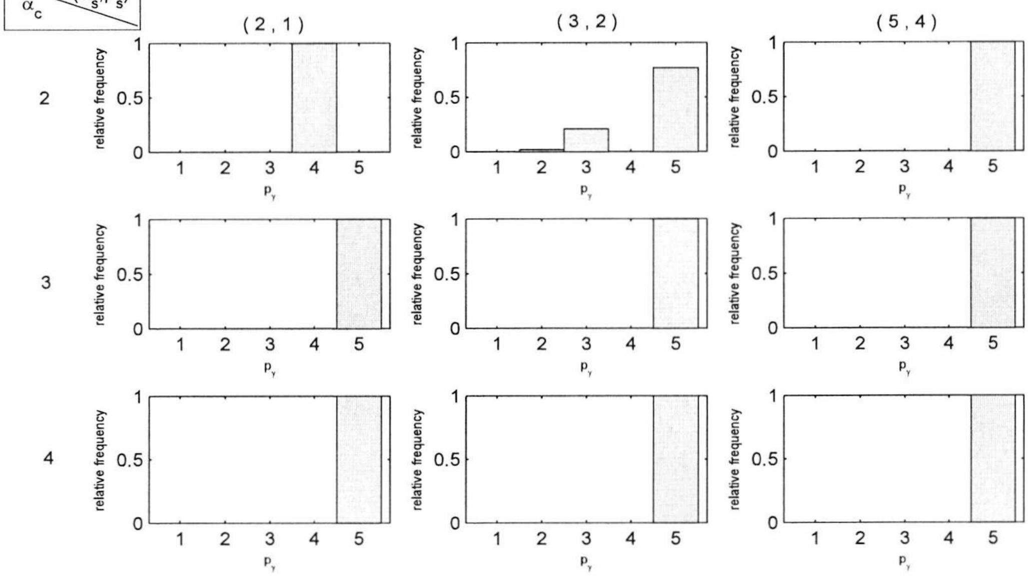

Figure 4.3.13: Cluster-Variable Heterogeneity with a prior on m for the principal components of the Crabs data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 1000$.

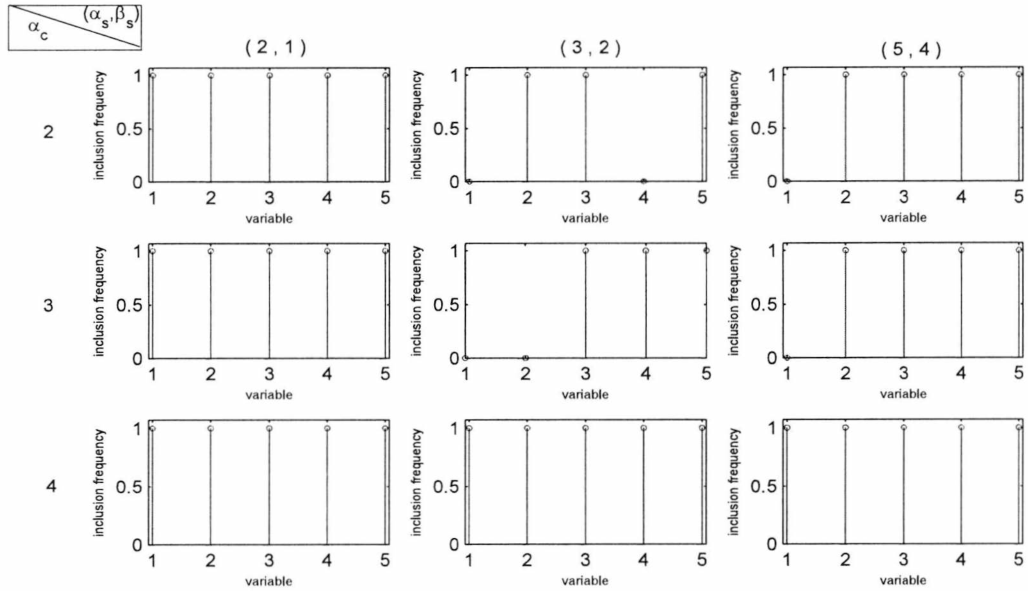

Figure 4.3.14: Cluster-Variable Heterogeneity with a prior on m for the principal components of the Crabs data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 100$.

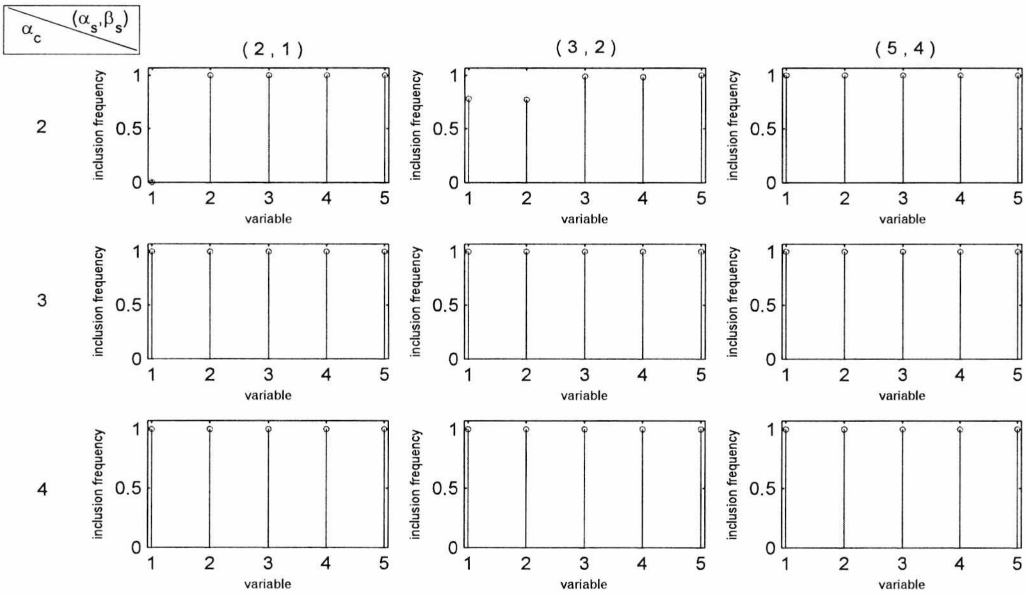

Figure 4.3.15: Cluster-Variable Heterogeneity with a prior on m for the principal components of the Crabs data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 1000$.

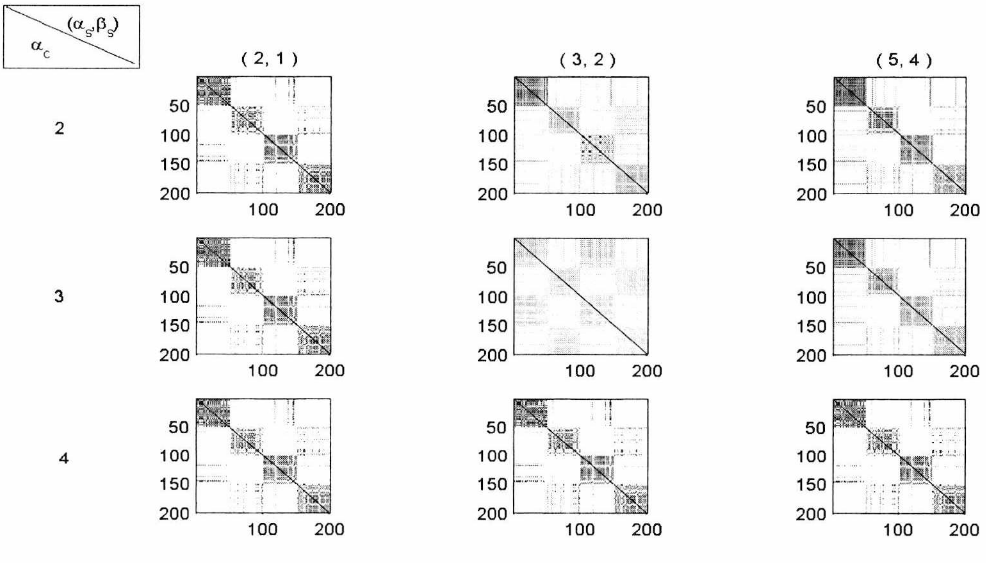

Figure 4.3.16: Cluster-Variable Heterogeneity with a prior on m for the principal components of the Crabs data set: Maps of the cluster allocations of the $n = 200$ observations for various combinations of (α_s, β_s) and α_c assuming heterogeneity and setting $h = 100$.

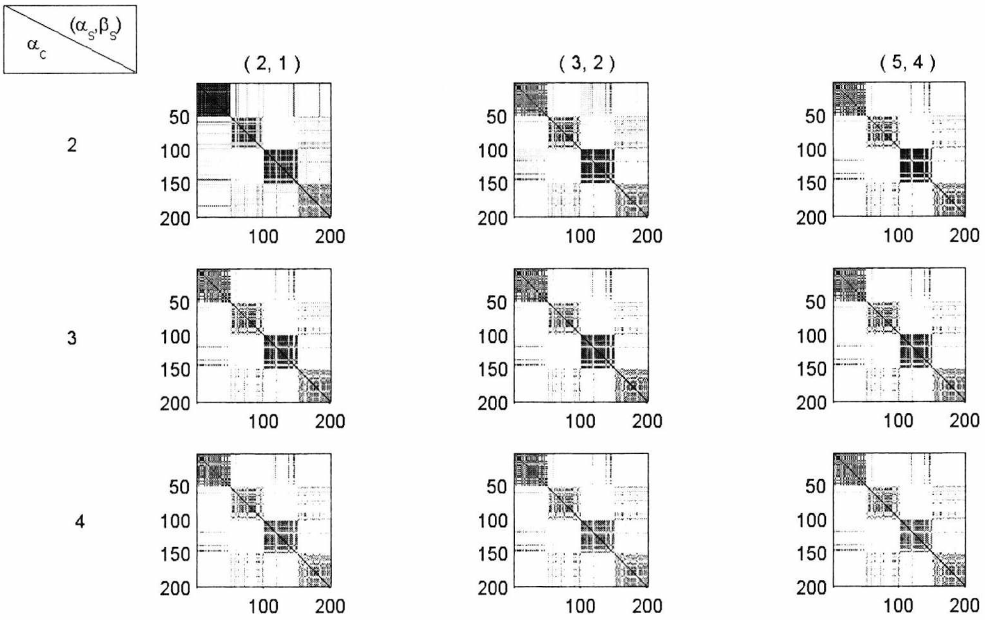

Figure 4.3.17: Cluster-Variable Heterogeneity with a prior on m for the principal components of the Crabs data set: Maps of the cluster allocations of the $n = 150$ observations for various combinations of (α_s, β_s) and α_c assuming heterogeneity and setting $h = 1000$.

Considering now the CVH model with a prior on h as well as on m and trying the pairs $(3, 200)$ and $(3, 400)$ as the hyperparameters of the IG prior on h, we observe that the results do not differ for the two setting. Therefore, results for the case of $h \sim IG(3, 400)$ can be found in the Appendix C (Figures C.0.20 - C.0.23).

We observe once again, that for most combinations of α_c and (α_s, β_s), all five principal components are selected with high probability of inclusion (Figures 4.3.18 and 4.3.19), while the resulting groupings (Figure 4.3.20) resemble those of the application of the CVH model with four clusters suggested and only a few observations misgrouped.

We should also note here that from the posterior distributions of h results that the estimated value of h, for all the cases we have tried, moves around the value of 10 (Figures 4.3.21).

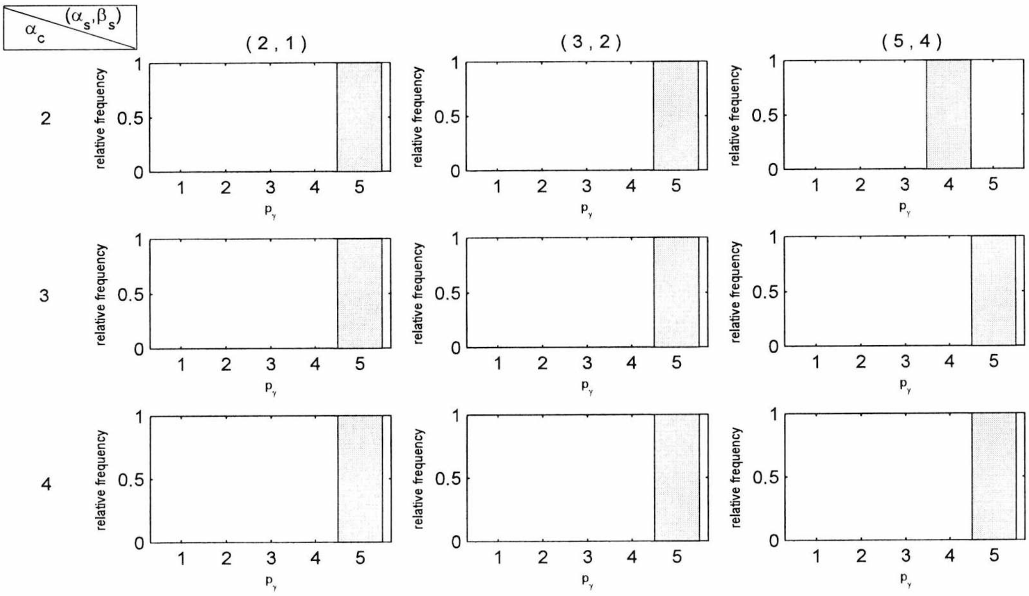

Figure 4.3.18: Cluster-Variable Heterogeneity with priors on m and h for the principal component of the Crabs data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

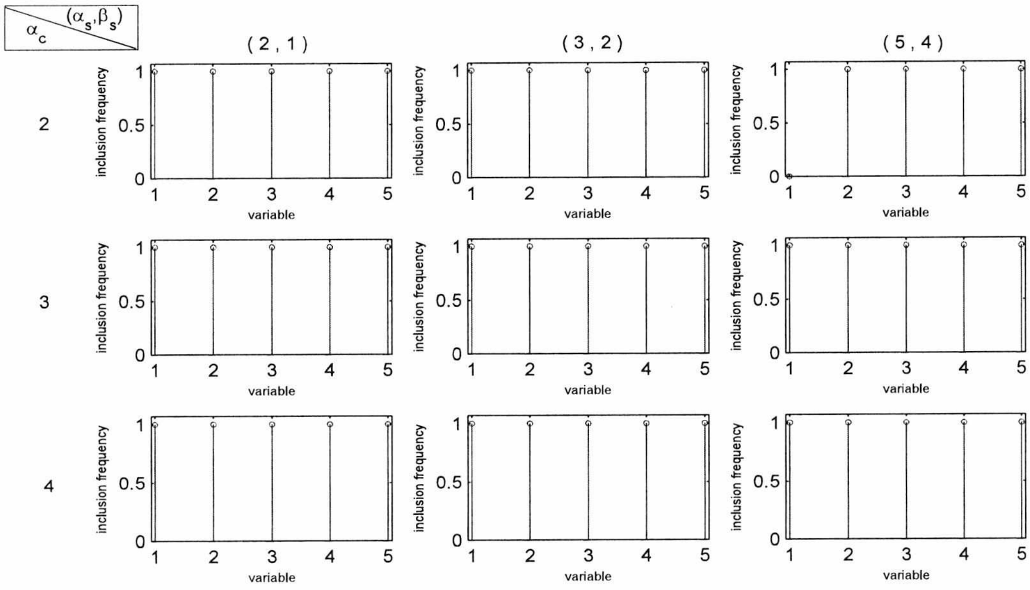

Figure 4.3.19: Cluster-Variable Heterogeneity with priors on m and h for the principal components of the Crabs data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

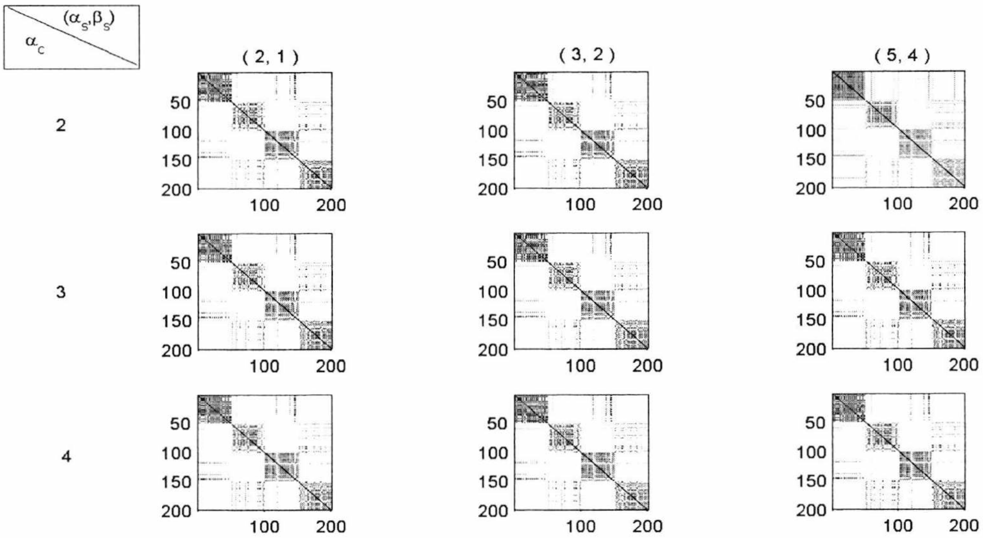

Figure 4.3.20: Cluster-Variable Heterogeneity with priors on m and h for the principal of the Crabs data set: Maps of the cluster allocations of the $n = 200$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

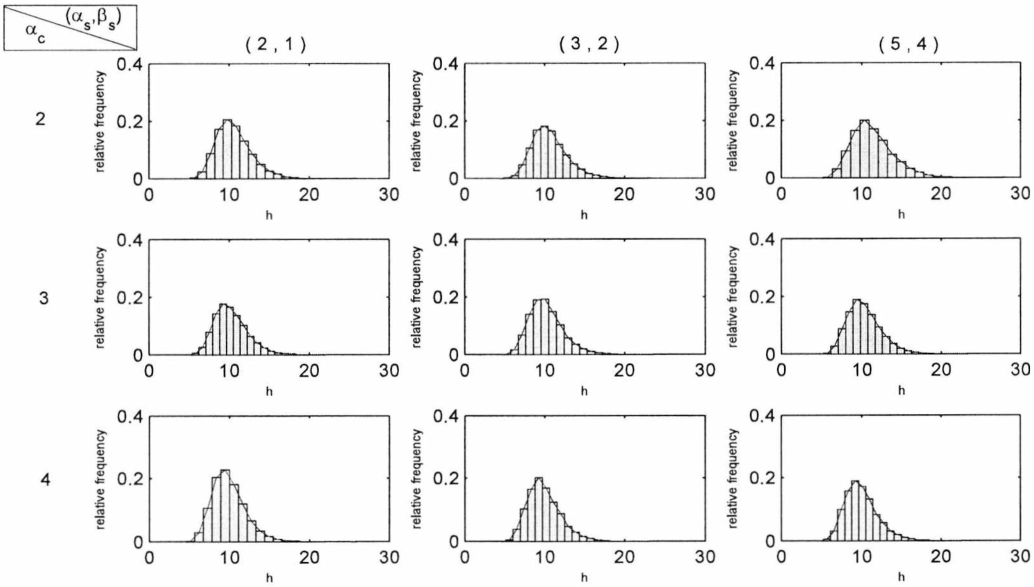

Figure 4.3.21: Cluster-Variable Heterogeneity with priors on m and h for the principal components of the Crabs data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

Finally, we have the results from the application of the non-conjugate case. For the different combinations of α_c and (α_s, β_s), we pick out three interesting cases. These are the combination of $\alpha_c = 2$ with $(\alpha_s, \beta_s) = (3, 2)$ and $(\alpha_s, \beta_s) = (5, 4)$ and that of $\alpha_c = 3$ with $(\alpha_s, \beta_s) = (5, 4)$. Examining the histograms for the number of selected principal components together with the probabilities of inclusion for each of the components (Figures 4.3.22 and 4.3.23), we note that it is under these cases that the non-conjugate model does not identify the full set of the five principal components as discriminating. Instead, setting $(\alpha_s, \beta_s) = (5, 4)$ and $\alpha_c = 2$ or $\alpha_c = 3$, the second principal component is excluded from the model, resulting to crabs being grouped with respect to their colour (that is orange and blue crabs) (Figure 4.3.24). Setting $(\alpha_s, \beta_s) = (5, 4)$ with $\alpha_c = 2$ on the other hand, suggests excluding the third principal component and leads to grouping observations of the same sex together.

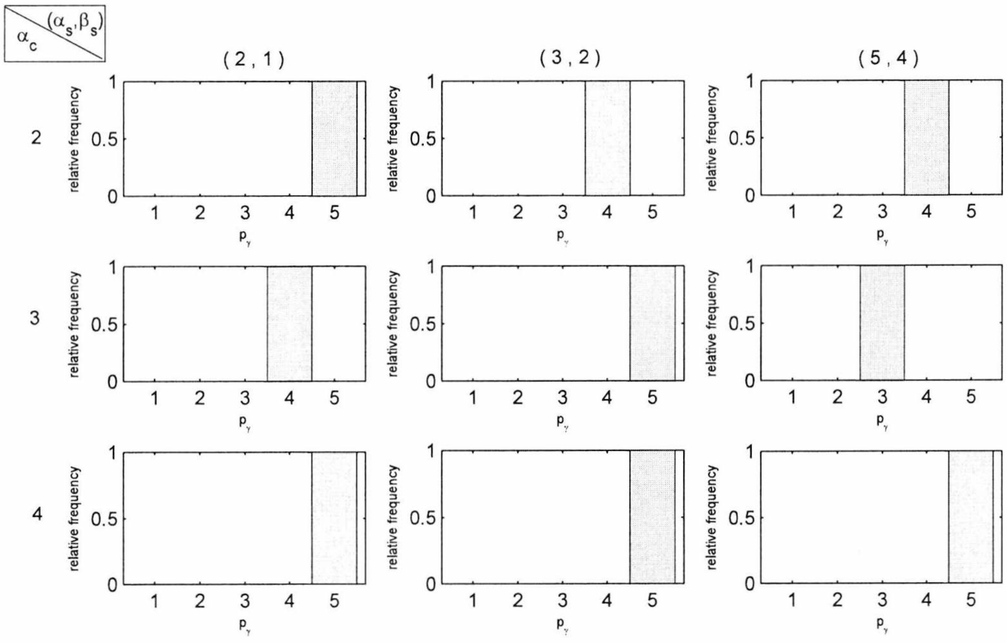

Figure 4.3.22: Non-Conjugate model for the principal components of the Crabs data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

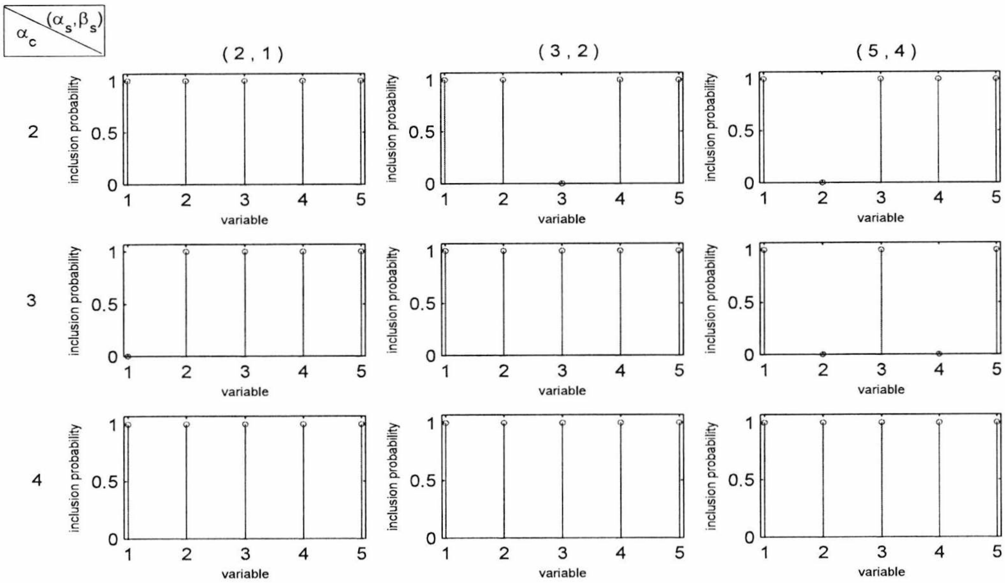

Figure 4.3.23: Non-Conjugate model for the principal components of th Crabs data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

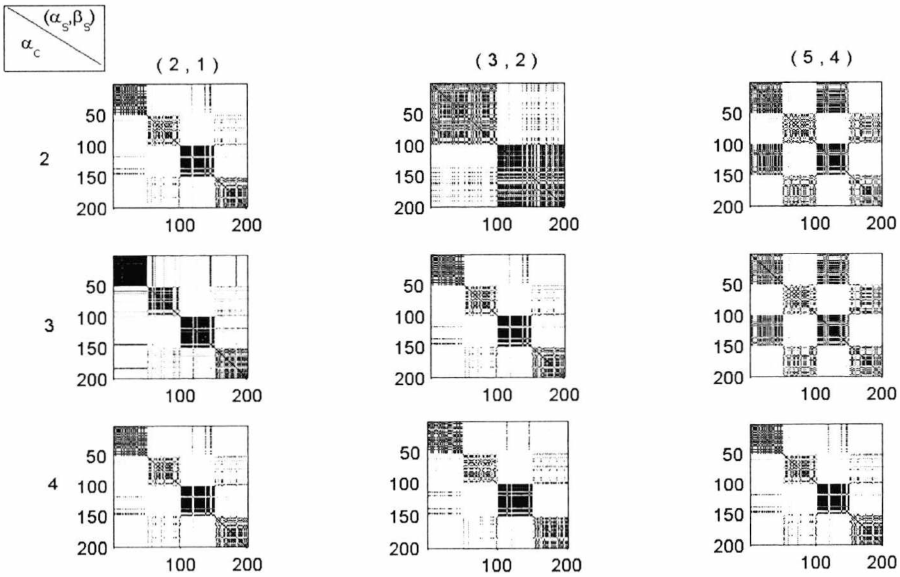

Figure 4.3.24: Non-Conjugate model for the principal components of the Crabs data set: Maps of the cluster allocations of the $n = 15$ observations for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

While all remaining cases favour the inclusion of all components in the analysis suggesting a reasonable structure of four clusters with a few missgrouped observations, we observe that $\alpha_c = 3$ with $(\alpha_s, \beta_s) = (2, 1)$ fully recovers the first and third clusters. Interestingly, the latter is with regard to the exclusion of the first principal component from the model.

To summarise, performing the models with the four different covariance structures presented in chapter 3 on the crabs data set and using the principal components obtained from the principal component analysis on the standardised data, we can first observe that the cluster structure of the crabs can be extracted from all methods except for the case of considering a covariance structure of the simplest form (CH model). The remaining three models, however, manage to identify the four clusters with just a few misgrouped observations, with most cases indicating that all five components are important for the analysis. Finally, there was no evidence of significant alterations on the results when under the consideration of different combinations of the various hyperparameters, i.e. $\alpha_c, \alpha_s, \beta_s, h$ etc, which appoints the robustness of the three models.

4.4 Arthritis data set

We will now apply the CVH method with priors on both h and m, as well as the Non-Conjugate model, on a real data set with the special characteristic of a limited sample size and a large number of variables, the arthritis data set. Arthritis is a joint disease causing degeneration of the human joints. Over 100 different forms of arthritis have been diagnosed depending on the aetiology and the pathogenesis of the disease. Among others, rheumatoid arthritis (RA) and osteoarthritis (OA) are two of the most common types of arthritis. Both types result in the erosion of two opposing bones, however, rheumatoid arthritis is an autoimmune condition that causes the body to attack its own soft-tissues and joints, while osteoarthritis is a condition usually caused by factors such as age, injury, and daily wear. The arthritis data set studies the discrimination between those two types of arthritis (RA and OA). To do so 755 genes have been measured on a total sample of 31 patients, 24 of which suffering from RA and the remaining 7 from OA (Sha et al., 2003).

Applying both the CVH method with priors on h, m and the Non-Conjugate model, the algorithms had difficulties in identifying the discriminating variables, with the proposed number of variables included in the model, p_{γ}, increasing along the chain. Convergence could not be achieved and therefore inference could not been drawn. Certainly, from the total number of 755 genes, the set of the discriminating ones is expected to be confined to only a small number of genes. As a result, we understand that the signal-to-noise ratio is particularly small. Therefore, along with the very few observation of the data set providing only little information, one could argue that the joint posterior of the non-discriminating genes obscures the signal of the discriminating ones. Under such circumstances, complications on the models applied for the entire set of the arthritis data are not surprising at all. Being interested in overcoming such abnormalities, we considered working with a smaller set of genes. On that direction, we looked at the work of Lamnisos et al. (2012). With regard to the identification of discriminating variables, Lamnisos et al. (2012) examine the prior choice of the coefficients of the probit regression model using a cross-validation criterion for the selection of the predictive covariates. Motivated by their results on their application on the arthritis data set and using the variables suggested for the case of $c = 1$ (Genes' ID : $20, 83, 145, 170, 225, 258, 290, 324, 332, 395, 473, 498, 665, 707, 728, 740, 742$), we applied the two models on a cut-down version of the arthritis data set, the results of which we give in the following figures. We used a chain of 100000 iterations with a burn-in period of 20000 iterations and thinning every 5 iterations. Using $G_{max} = 5$, p_{prior} with $a + b = 2$, and $a_u = 2$, $a_w = 1$, we investigated the cases of $\alpha_c = 2, 3, 4$ and $(\alpha_s, \beta_s) = (2,1), (3,2), (5,4)$.

Starting with the CVH model with priors on h and m, we tried an IG prior on h with parameters $(3, 200)$ and $(3, 400)$. Both cases (Figures 4.4.1 - 4.4.3 and C.0.24 - C.0.26 respectively) gave us similar results. From the histograms for the number of total number of selected variables p_γ (Figures 4.4.1 and C.0.24) we can see that for all cases, all variables are included in the model with high probabilities of inclusion (Figures 4.4.2 and C.0.25). As far as the suggested clustering is concerned, looking at the clustering maps in Figures 4.4.3 and C.0.26 we observe that a total number of three clusters is suggested throughout all different combinations. While patients $1, 3, 4, 5$ and 7 form a single group, the remaining patients are split in two further groups.

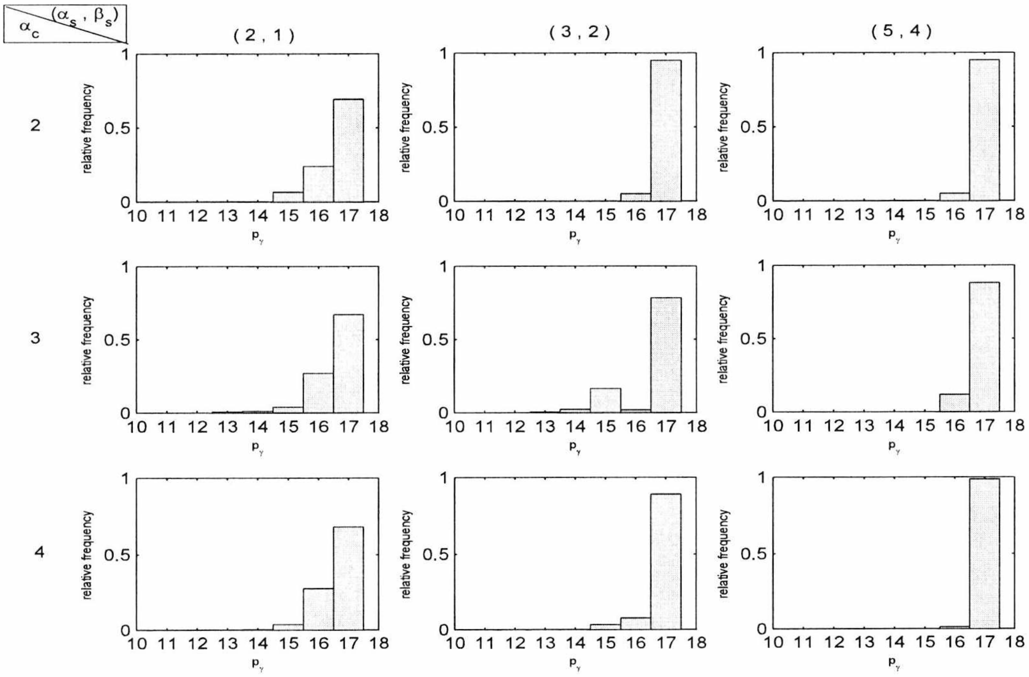

Figure 4.4.1: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

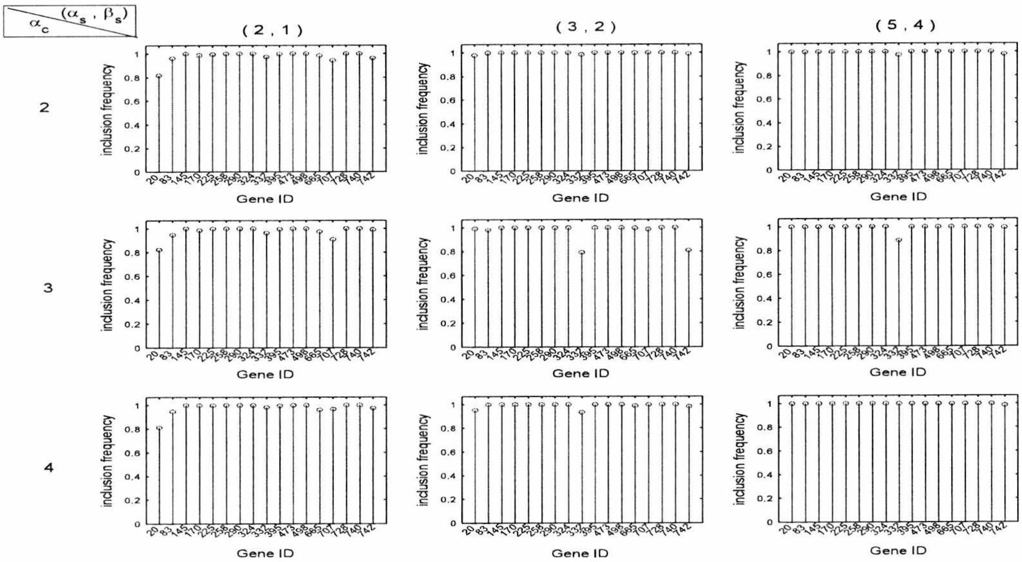

Figure 4.4.2: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

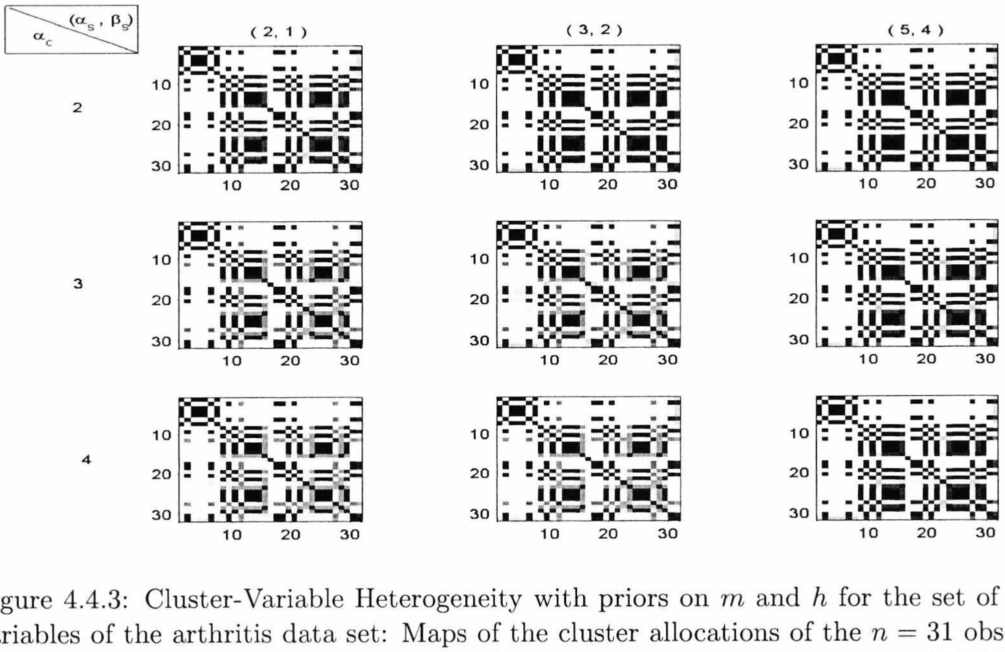

Figure 4.4.3: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Maps of the cluster allocations of the $n = 31$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

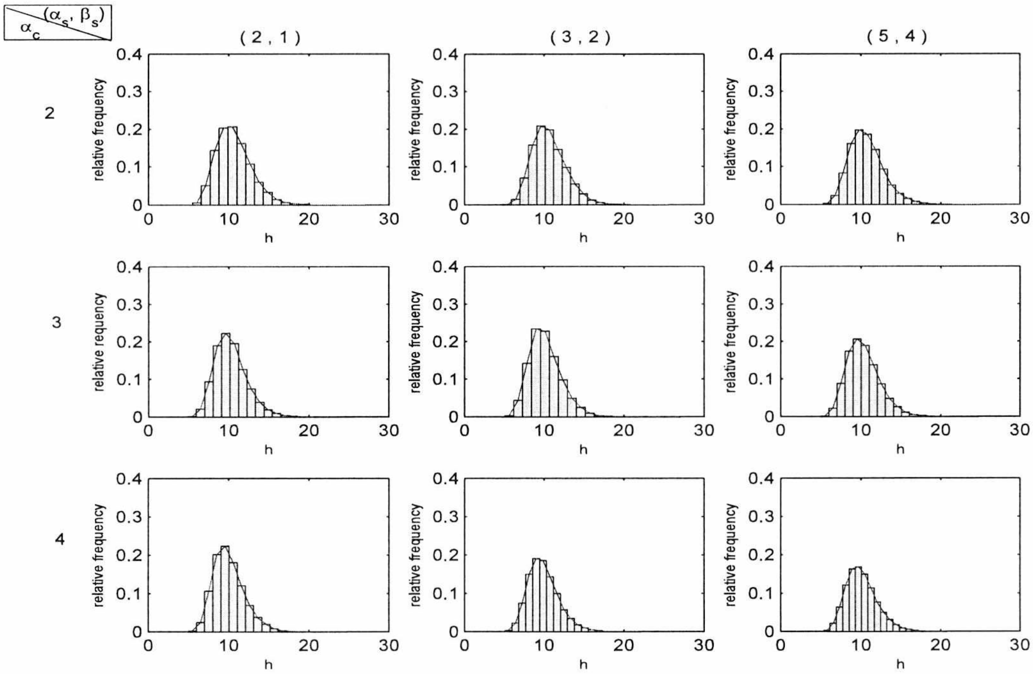

Figure 4.4.4: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

Knowing that our observations form two clusters (rheumatoid arthritis and osteoarthritis), with the first 7 observations forming one group of patients and the remaining 24 the second one, one could assume that the third group of patients is formed based on an additional common characteristic, e.g. sex or age group; such information, however, is unfortunately not available.

Interestingly, under the Non-Conjugate case, the resulting histograms of the total number of discriminating variables and the plots with the inclusion probabilities in Figures 4.4.5 and 4.4.6 respectively, suggest the inclusion of all 17 variables used in the analysis. In the meantime, the proposed grouping extracted from the clustering maps of Figure 4.4.7 resembles the formulation of the 3 groups as suggested by the CVH model.

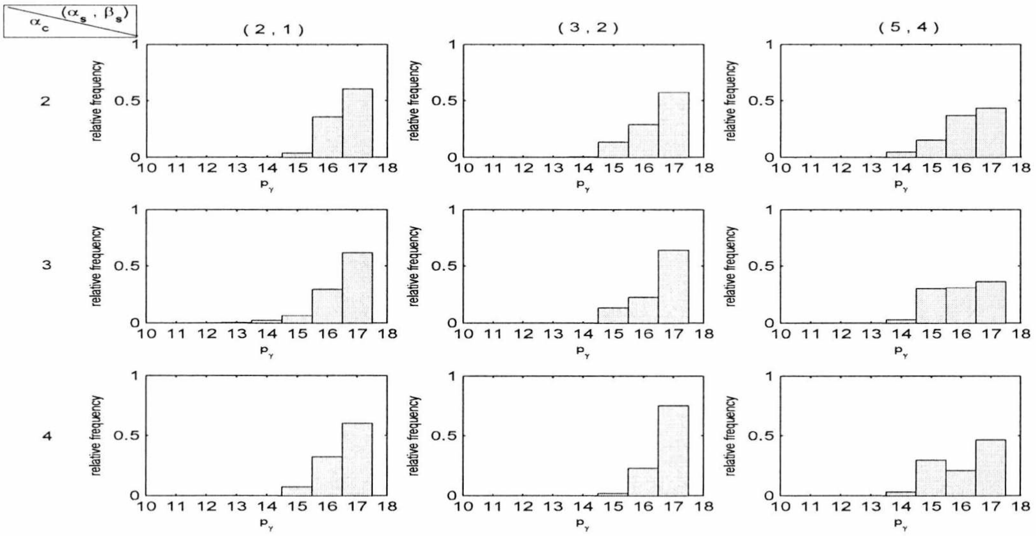

Figure 4.4.5: Non-Conjugate model for the set of 17 variables of the arthritis data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

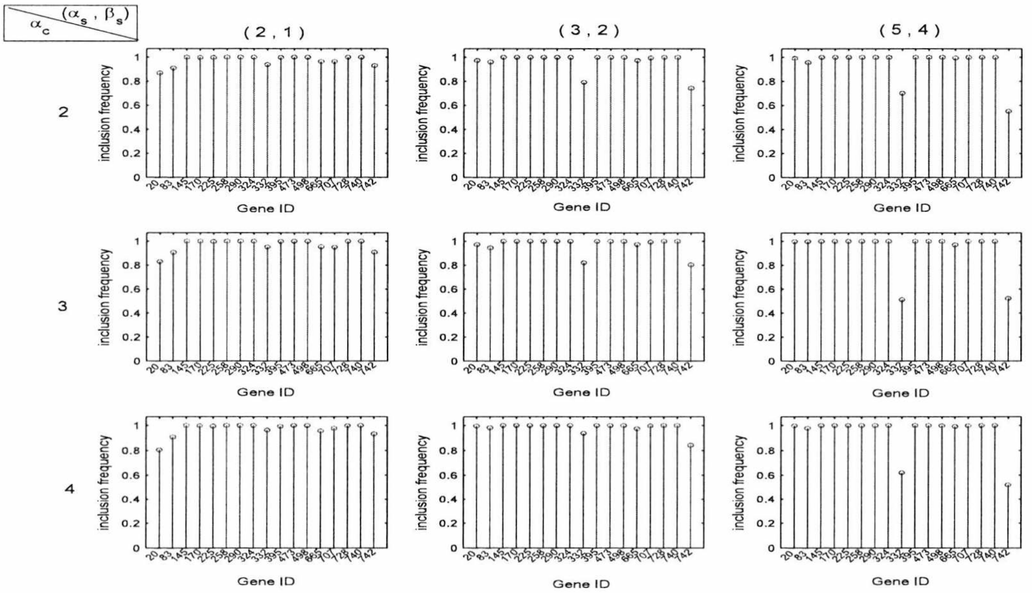

Figure 4.4.6: Non-Conjugate model for the set of 17 variables of the arthritis data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

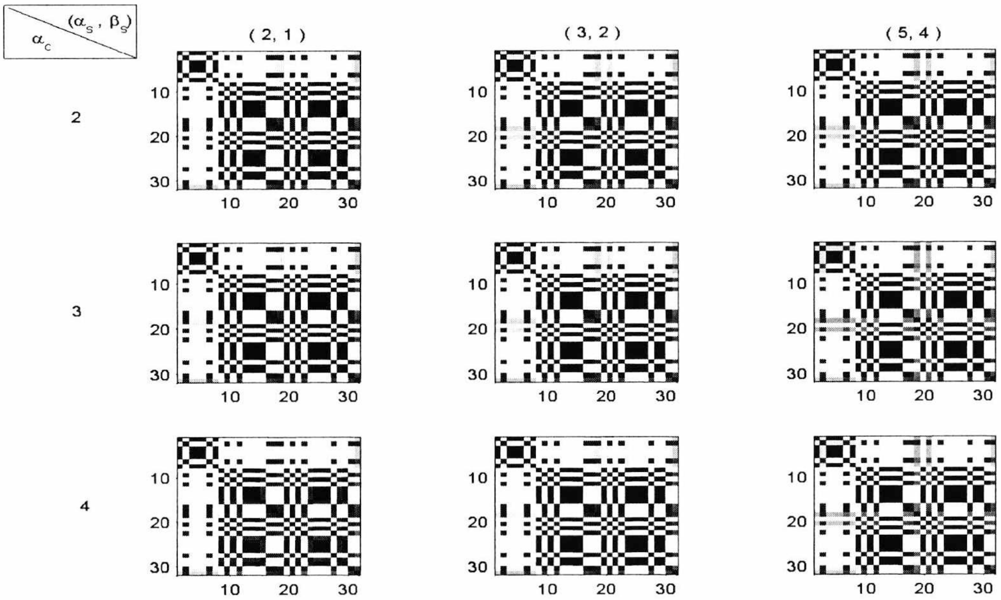

Figure 4.4.7: Non-Conjugate model for the set of 17 variables of the arthritis data set: Maps of the cluster allocations of the $n = 31$ observations for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

Finally, once, as stated at the beginning of this section, we do not use the original set of the 755 genes, we would like to examine the behaviour of our models for the set of the 17 selected variables augmented by 10 "junk" variables. Adding "junk" information in our data set, we are interested in investigating whether the developed models can identify those 10 variables as non-discriminating and exclude them from the analysis. Such exclusion would indicate that our models work on the right path. Therefore, we added 10 variables randomly sampled from a multivariate Normal with mean 0 and covariance matrix I and applied the two models (CVH with priors on h, m and Non-Conjugate) with the same hyperparameters' settings. Looking at the histograms and the posterior inclusion probabilities of the now 27 variables for the CVH model (Figures 4.4.8, 4.4.9 and 4.4.12, 4.4.9 respectively), under all cases, the 10 junk variables are indicated as non-important. It is interesting though, that under the case of $(\alpha_s, \beta_s) = (2, 1)$ and regardless h, the variables selected are only 10. Also, under all cases, genes

145, 170, 225, 258, 290, 324, 473, 498, 728 and 740 are the ones always having high probability of inclusion.

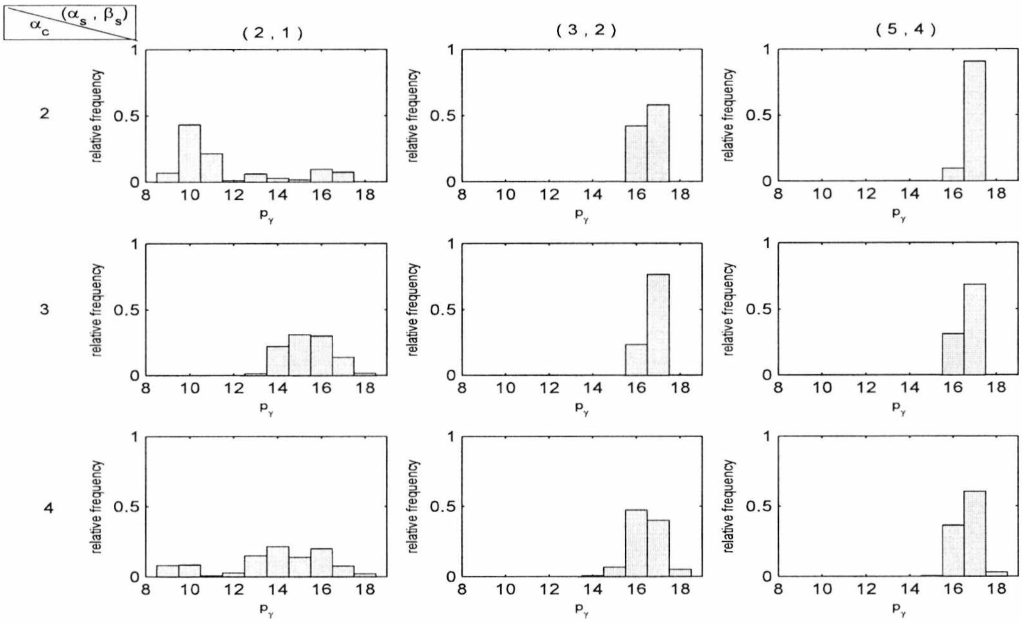

Figure 4.4.8: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 200)$, with unequal covariance matrices.

Coming to the application of the Non-Conjugate model, we can now see that the number of selected variables indicated in Figure 4.4.16 varies a lot. However, looking at the probabilities of inclusion (Figure 4.4.17), we observe that alike the CVH model genes 145, 170, 225, 258, 290, 324, 473, 498, 728 and 740 are always considered within the discriminating ones.

Although the selected discriminating genes may vary throughout the two models and for the various different combinations of α_c, (α_s, β_s), as well as the hyperparameters (α_h, β_h), the suggested clustering is always the same. Throughout the clustering maps of the CVH model in Figures 4.4.10 and 4.4.14 and those for the Non-Conjugate model (Figure 4.4.18), we see that all cases suggest 3 clusters that actually resemble the groups proposed when considering the set of 17 genes.

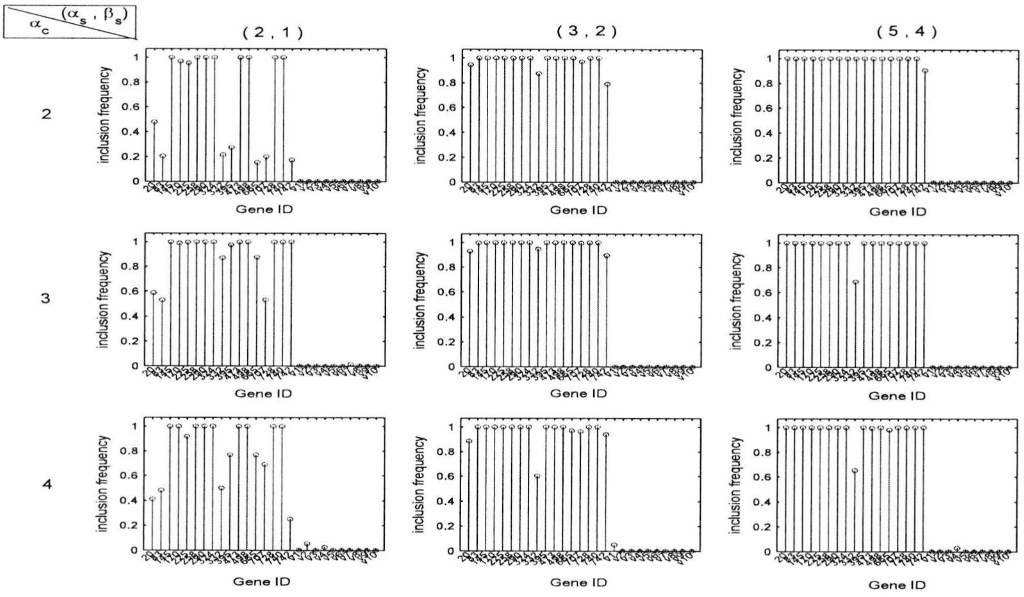

Figure 4.4.9: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

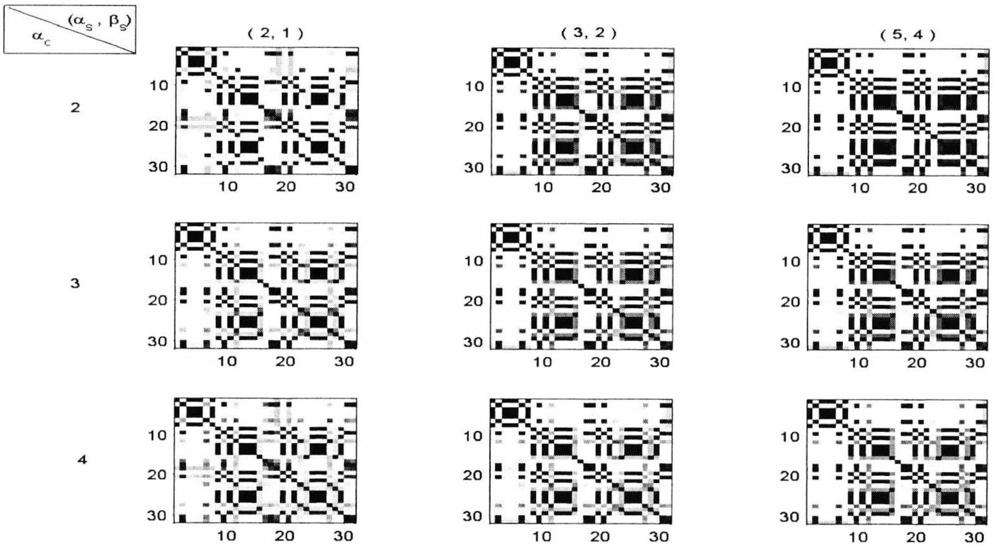

Figure 4.4.10: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Maps of the cluster allocations of the $n = 31$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

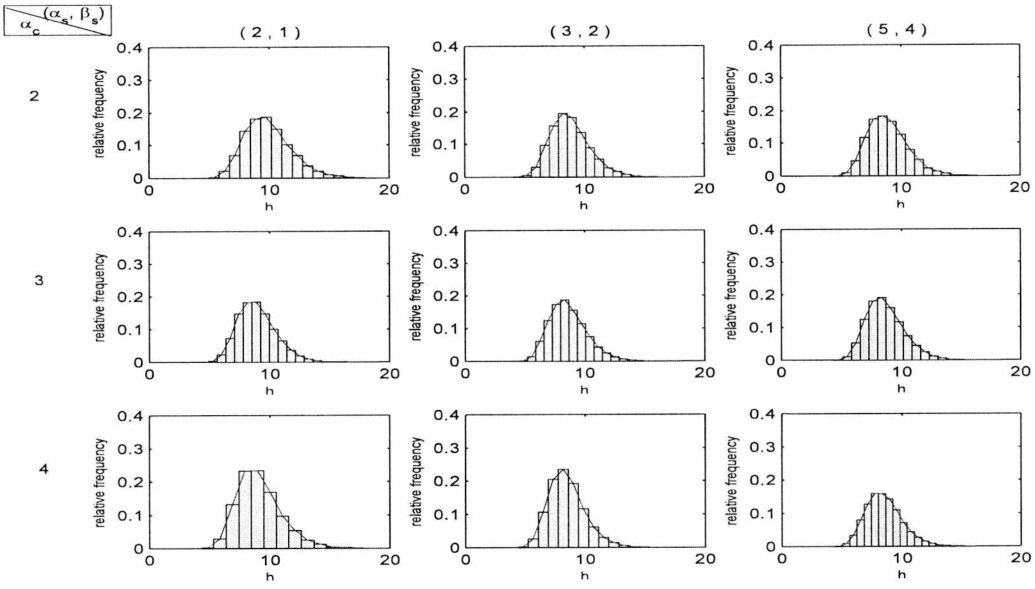

Figure 4.4.11: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 200)$, with unequal covariance matrices.

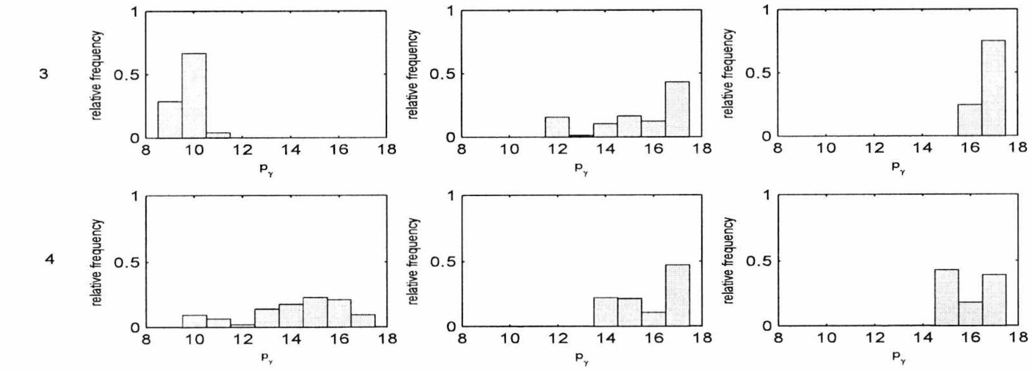

Figure 4.4.12: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 400)$, with unequal covariance matrices.

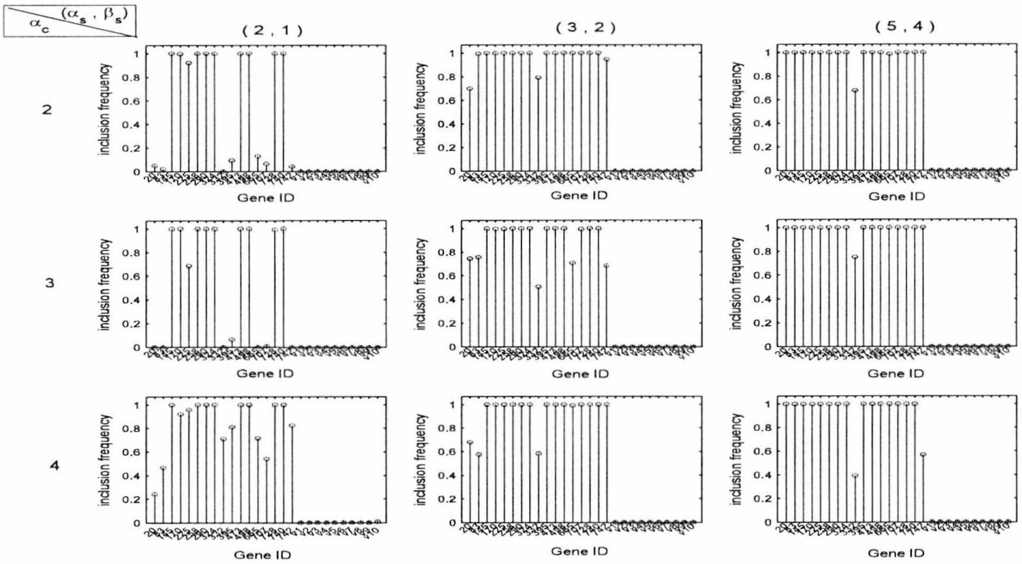

Figure 4.4.13: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 400)$, with unequal covariance matrices.

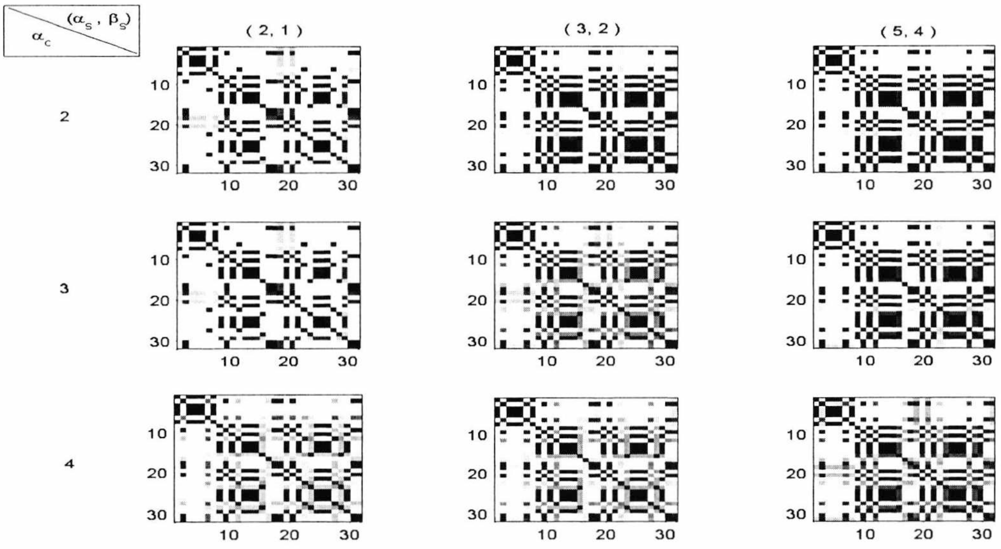

Figure 4.4.14: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Maps of the cluster allocations of the $n = 31$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG\,(3, 400)$, with unequal covariance matrices.

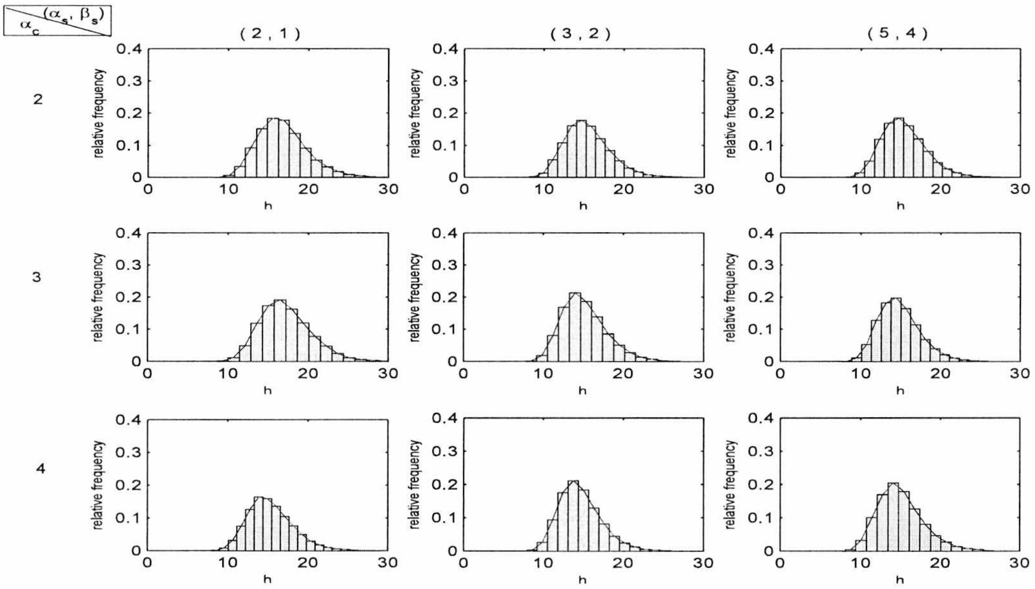

Figure 4.4.15: Cluster-Variable Heterogeneity with priors on m and h for the set of 27 variables of the arthritis data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

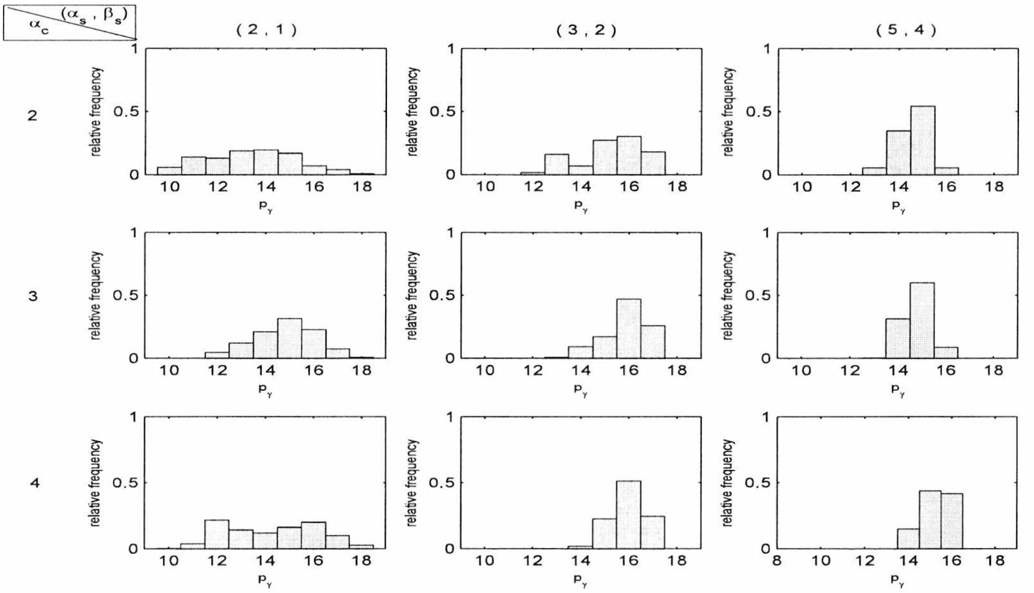

Figure 4.4.16: Non-Conjugate model for the set of 27 variables of the arthritis data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

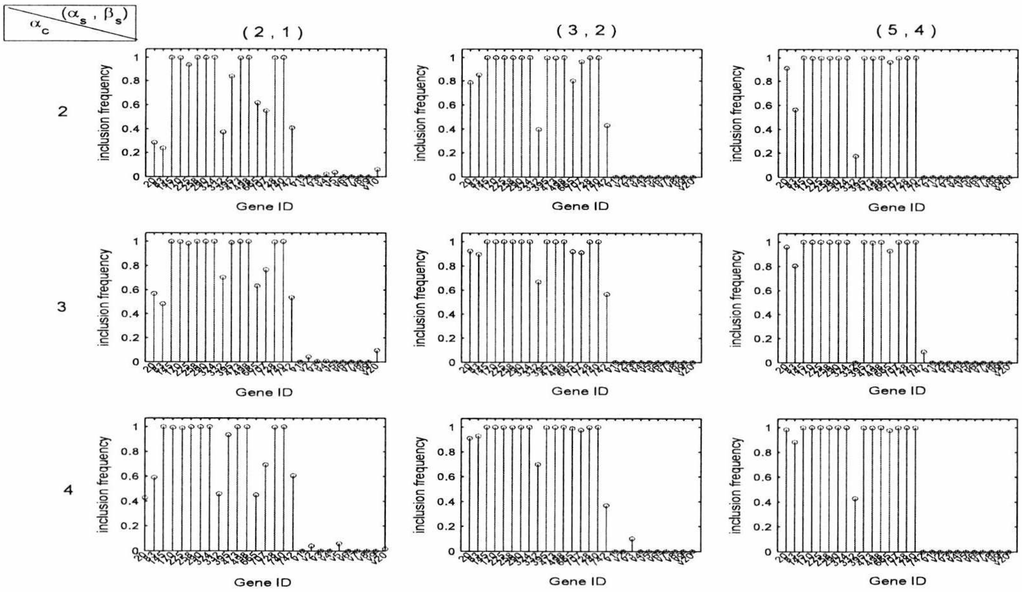

Figure 4.4.17: Non-Conjugate model for the set of 27 variables of the arthritis data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

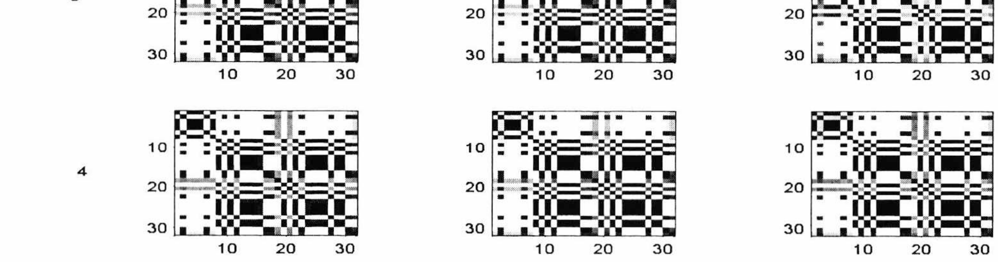

Figure 4.4.18: Non-Conjugate model for the set of 27 variables of the arthritis data set: Maps of the cluster allocations of the $n = 31$ observations for various combinations of (α_s, β_s) and α_c, assuming unequal covariance matrices.

4.5 Conclusions

We presented applications of the different models introduced in this thesis on three sets of real data; the iris, the crabs and the arthritis data sets. We shall now conclude, summarising and comparing our findings, in view of a complete and clear understanding of the methodology developed in this thesis.

Starting with the Iris data set and with known the problem of the two overlapping *Iris versicolor* and *Iris virginica* species, we met a difficulty in the recovery of the cluster structure under the two computational methods suggested in chapter 2. With the algorithms having problems in identifying any sort of cluster structure, we moved on the application of the CH model with equally discouraging results for most of the cases under examination. The CVH with a prior on m, CVH with priors on m, h, and the Non-Conjugate models though, indicated a considerable improvement on the clustering procedure. For all but two cases under the Non-Conjugate model, results suggested the inclusion of all four variables and gave resembling cluster structures that met our expectations of a few missgrouped observations. For the two exceptional cases of the Non-Conjugate model however, with only three variables being included in the model, the algorithm manages to fully recover the *Iris setosa* group, with the problem of overlapped *Iris versicolor* and *Iris virginica* still present.

On the principal components of the crabs data set, methods of chapter 2 fail in the suggestion of a selected number of components, while the four clusters cannot be identified. Suggested cluster structures of the CH model are also disappointing. Once again though, models : CVH with a prior on m, CVH with priors on m, h, as well as the Non-Conjugate model perform particularly well. With most cases suggesting all five principal components as important, we have structures of the four original clusters (blue male, orange male, blue female, orange female), with of course a number of observations being missgrouped.

Finally, we examined two cut down versions of the arthritis data set; one with 17 variables as suggested by Lamnisos et al. (2012), and another with an additional set of 10, randomly generated from a $N(0, I)$, noisy variables. This time we concentrated only on the CVH with priors on m, h and the Non-Conjugate models. As far as the set of 17 variables is concerned, both models propose the inclusion of all variables, suggesting resembling cluster structures. Looking at the set of the 27 variables on the other hand, we have a small variation on the number of selected variables. Surprisingly, under a few cases, less variables are selected as discriminating. Under all circumstances though, the 10 noisy variables are excluded from the model. Interestingly enough, for both models and for all cases, regardless the number of selected variables, the suggested clusters are similar and close to those suggested under the consideration of the 17 variables.

All in all and taking into account the results of all models' applications on our simulated data set, we can conclude that the best performing models, on both the tasks of variable selection and clustering, are the Cluster-Variable Heterogeneity model with priors on the mean value m and the hyperparameter h, together with the Non-Conjugate model.

Chapter 5

Conclusions and Future Work

With the vast development of technology over the past few years offering easy access to an increased amount of information, we saw in this thesis how a numerous amount of disciplines such as medicine, biology, etc., are in need of methods that can manipulate data sets of high dimensionality, yet of a limited sample size. The idea of reducing the dimensionality of the problem and its importance when combined with the clustering methodology has been therefore introduced. With the cluster structure of the data being usually confined to a small subset of variables, the identification of the important variables, using variable selection methods or by differentially weighting the available variables, is of interest. Using the chosen subset of the discriminating variables, the true cluster structure can be therefore recovered.

Focusing on the Bayesian methodology and being interested in performing the tasks of variable selection and clustering simultaneously, we considered the use of a latent vector for the selection of the discriminating variables, while finite mixture models were used to form the clustering procedure. Markov Chain Monte Carlo chains were used to draw inference on the parameters of interest, while an unknown number of components has been considered, estimation on which is performed via the Reversible Jump MCMC sampler. Motivated by the work of Tadesse et al. (2005), we examined different computational approaches for the performance of the RJMCMC sampler. Suggesting a split/merge move on either empty or non-empty components and proposing splits using a random variable from a Beta distribution, we observed a fair sensitivity on the hyperparameters settings associated to the prior formulation of the covariance structure of the model. Application on a simulated data set showed a problem on the splitting of the originally well separated groups. Back to the choice of the crucial hyperparameters and relying on a rather intuitive selection, we chose to apply an Empirical Bayes approach for the determination of their values. Further into investigating the problem of suggesting proper splits, the conjugate case of the SAMS sampler of Dahl (2005), proposing allocations conditioned on the previously allocated observations, was also considered. The sensitivity on the choice of the hyperparameters of the covariance structure was still present though.

Naturally, in search of ways to overcome the observed sensitivity, we considered exploring the covariance structure of the model. Starting by forming a model with covariance matrices in their simplest form, i.e. proportional to the Identity matrix, we built up five different models setting priors on various hyperparameters, the choice of which seemed crucial for the selection of the important variables and the performance of the clustering procedure. We also considered the model with prior settings of the non-conjugate form. For all five suggested structures we examined their performance on a simulated data set. While initially, for the CH, CVH and CVH with a hyperprior on the mean value m models, the impact of the hyperparameter h was still present, affecting mixing and convergence to the desired cluster formulation, setting a hyperprior on h seemed a reasonable choice. Allowing the algorithm to sample values of h, convergence to the correct set of discriminating variables was achieved leading to allocations of the correct form. Similar encouraging results were obtained under the non-conjugate formulation.

Finishing this thesis, we applied the various methods presented on three different data sets : the well-known iris data, as well as the crabs and the arthritis data sets. With particular difficulties met under the consideration of the structure suggested by Tadesse et al. (2005), the CVH with priors on h and mean m model, as well as the Non-Conjugate model performed reasonably well; considering of course the particular features of the data under study, e.g the well-known overlapped *Iris versicolor* and *Iris virginica* species of the iris data set.

On future work on the subject, further exploration on the covariance structures is a natural path we shall follow. Having concentrated on the case of covariance matrices of no correlation, extending the work into considering correlated variables is of particular interest although possibly demanding. Studying intra-class correlation, on the other hand, could also be of interest.

Finally, working in higher dimensions is known to be remarkably challenging; recall the arthritis data set and the difficulties met applying our methodology in the entire set of genes. With a large number of iterations needed to achieve convergence and the usually very many hyperparameters one has to experiment with, adaptive Monte Carlo methods which automatically tune the parameters of the proposal distribution suggesting good estimates of posterior summaries can be proved to be of great use in variable selection problems. Adaptive MCMC methods with applications on variable selection problems within the regression context have actually been recently developed [Ji and Schmidler (2009), Lamnisos et al. (2011)]. Potential extension to our models presented in this thesis could be considered to improve mixing.

Appendix A

Appendix for Chapter 2

Calculations on the joint posterior $f\left(X, y|G, w, \gamma\right)$ considering Homogeneous Covariances :

$$
\begin{aligned}
f\left(X, y|G, w, \gamma\right) &= \int f\left(X, y|\gamma, G, w, \mu^{D}, \Sigma^{SD}, \mu^{ND}, \Sigma^{ND}\right) f\left(\mu^{D}|G, \gamma, \Sigma^{D}\right) f\left(\Sigma^{D}|G, \gamma\right) \\
&\quad \times f\left(\mu^{ND}|\Sigma^{ND}, \gamma\right) f\left(\Sigma^{ND}|\gamma\right) d\mu^{D}\, d\Sigma^{D}\, d\mu^{ND}\, d\Sigma^{ND} \\
&= \int \prod_{k=1}^{G} \left[w_k^{n_k} \left(2\pi\right)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} \left|\Sigma^{D}\right|^{-\frac{(n_k+1)}{2}} \right] \\
&\quad \times \exp\left[-\frac{1}{2} \sum_{k=1}^{G} \sum_{x_i \in k} \left(x_i^{D} - \mu_k^{D}\right)' \Sigma^{D^{-1}} \left(x_i^{D} - \mu_k^{D}\right) \right] \\
&\quad \times \exp\left[-\frac{1}{2} \sum_{k=1}^{G} \left(\mu_k^{D} - \mu_0^{D}\right)' \left(h_1 \Sigma^{D}\right)^{-1} \left(\mu_k^{D} - \mu_0^{D}\right) \right] \\
&\quad \times 2^{-\frac{p_\gamma(\delta+p_\gamma-1)}{2}} \pi^{-\frac{p_\gamma(p_\gamma-1)}{4}} \left|Q^{D}\right|^{\frac{(\delta+p_\gamma-1)}{2}} \left|\Sigma^{D}\right|^{-\frac{(\delta+2p_\gamma)}{2}} \\
&\quad \times \left[\prod_{j=1}^{p_\gamma} \Gamma\left(\frac{\delta + p_\gamma - j}{2}\right) \right]^{-1} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{D^{-1}} Q^{D}\right) \right] \\
&\quad \times \left(2\pi\right)^{-\frac{(n+1)(p-p_\gamma)}{2}} h_0^{-\frac{(p-p_\gamma)}{2}} \left|\Sigma^{ND}\right|^{-\frac{(n+1)}{2}} \\
&\quad \times \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(x_i^{ND} - \mu^{ND}\right)' \Sigma^{ND^{-1}} \left(x_i^{ND} - \mu^{ND}\right) \right]
\end{aligned}
$$

$$
\begin{aligned}
&\times \exp\left[-\frac{1}{2}\left(\mu^{ND}-\mu_0^{ND}\right)'\left(h_0\Sigma^{ND}\right)^{-1}\left(\mu^{ND}-\mu_0^{ND}\right)\right]\\
&\times 2^{-\frac{(p-p_\gamma)(\delta+p-p_\gamma-1)}{2}}\pi^{-\frac{(p-p_\gamma)(p-p_\gamma-1)}{4}}\left[\prod_{j=1}^{p-p_\gamma}\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right]^{-1}\\
&\times \left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}}\left|\Sigma^{ND}\right|^{-\frac{[\delta+2(p-p_\gamma)]}{2}}\\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}}Q^{ND}\right)\right]d\mu^D\,d\Sigma^D\,d\mu^{ND}\,d\Sigma^{ND}\\
={}&\int\prod_{k=1}^{G}\left\{w_k^{n_k}\left(2\pi\right)^{-\frac{(n_k+1)p_\gamma}{2}}h_1^{-\frac{p_\gamma}{2}}\left|\Sigma^D\right|^{-\frac{(n_k+1)}{2}}\right.\\
&\left.\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{D^{-1}}S_k^D\right)\right]\right\}\\
&\times \exp\left\{-\frac{1}{2}\sum_{k=1}^{G}\left(\mu_k^D-\frac{\sum_{x_i\in k}x_i^D+h_1^{-1}\mu_0^D}{n_k+h_1^{-1}}\right)'\left(n_k+h_1^{-1}\right)\Sigma^{D^{-1}}\right.\\
&\left.\times\left(\mu_k^D-\frac{\sum_{x_i\in k}x_i^D+h_1^{-1}\mu_0^D}{n_k+h_1^{-1}}\right)\right\}\\
&\times 2^{-\frac{p_\gamma(\delta+p_\gamma-1)}{2}}\pi^{-\frac{p_\gamma(p_\gamma-1)}{4}}\left|Q^D\right|^{\frac{(\delta+p_\gamma-1)}{2}}\left|\Sigma^D\right|^{-\frac{(\delta+2p_\gamma)}{2}}\\
&\times\left[\prod_{j=1}^{p_\gamma}\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\right]^{-1}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{D^{-1}}Q^D\right)\right]\\
&\times\left(2\pi\right)^{-\frac{(n+1)(p-p_\gamma)}{2}}h_0^{-\frac{(p-p_\gamma)}{2}}\left|\Sigma^{ND}\right|^{-\frac{(n+1)}{2}}\\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}}S^{ND}\right)\right]\\
&\times \exp\left\{-\frac{1}{2}\left(\mu^{ND}-\frac{\sum_{i=1}^{n}x_i^{ND}+h_0^{-1}\mu_0^{ND}}{n+h_0^{-1}}\right)'\left(n+h_0^{-1}\right)\Sigma^{ND^{-1}}\right.\\
&\left.\times\left(\mu^{ND}-\frac{\sum_{i=1}^{n}x_i^{ND}+h_0^{-1}\mu_0^{ND}}{n+h_0^{-1}}\right)\right\}\\
&\times 2^{-\frac{(p-p_\gamma)(\delta+p-p_\gamma-1)}{2}}\pi^{-\frac{(p-p_\gamma)(p-p_\gamma-1)}{4}}\\
&\times\left[\prod_{j=1}^{p-p_\gamma}\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right]^{-1}\\
&\times \left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}}\left|\Sigma^{ND}\right|^{-\frac{[\delta+2(p-p_\gamma)]}{2}}\\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}}Q^{ND}\right)\right]d\mu^D\,d\Sigma^D\,d\mu^{ND}\,d\Sigma^{ND}
\end{aligned}
$$

$$
\begin{aligned}
&= (2\pi)^{-\frac{np}{2}} \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-\frac{p_\gamma}{2}} \right] \left|Q^D\right|^{\frac{(\delta+p_\gamma-1)}{2}} \\
&\quad \times \left[\prod_{j=1}^{p_\gamma} \Gamma\left(\frac{\delta + p_\gamma - j}{2} \right) \right]^{-1} \\
&\quad \times 2^{-\frac{p_\gamma(\delta+p_\gamma-1)}{2}} \pi^{-\frac{p_\gamma(p_\gamma-1)}{4}} 2^{-\frac{(p-p_\gamma)(\delta+p-p_\gamma-1)}{2}} \pi^{-\frac{(p-p_\gamma)(p-p_\gamma-1)}{4}} \\
&\quad \times \left(h_0 n + 1\right)^{-\frac{(p-p_\gamma)}{2}} \left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}} \\
&\quad \times \left[\prod_{j=1}^{p-p_\gamma} \Gamma\left(\frac{\delta + p - p_\gamma - j}{2} \right) \right]^{-1} \\
&\quad \times \int \left|\Sigma^D\right|^{\frac{-(n+\delta+2p_\gamma)}{2}} \left|\Sigma^{ND}\right|^{-\frac{[n+\delta+2(p-p_\gamma)]}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\Sigma^{D^{-1}} \left[Q^D + \sum_{k=1}^{G} S_k^D \right] \right) \right\} \\
&\quad \times \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}} \left[Q^{ND} + S^{ND} \right] \right) \right\} d\Sigma^D \, d\Sigma^{ND} \\
&= \pi^{\frac{-np}{2}} \left|Q^D\right|^{\frac{(\delta+p_\gamma-1)}{2}} \left| Q^D + \sum_{k=1}^{G} S_k^D \right|^{\frac{-(n+\delta+p_\gamma-1)}{2}} \\
&\quad \times \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-\frac{p_\gamma}{2}} \right] \prod_{j=1}^{p_\gamma} \frac{\Gamma\left(\frac{1}{2}\left(n + \delta + p_\gamma - j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta + p_\gamma - j\right)\right)} \\
&\quad \times \left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}} \left|Q^{ND} + S^{ND}\right|^{\frac{-(n+\delta+p-p_\gamma-1)}{2}} \\
&\quad \times \left(h_0 n + 1\right)^{-\frac{(p-p_\gamma)}{2}} \prod_{j=1}^{p-p_\gamma} \frac{\Gamma\left(\frac{1}{2}\left(n + \delta + p - p_\gamma - j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta + p - p_\gamma - j\right)\right)}, \qquad \text{(A.1)}
\end{aligned}
$$

where S_k^D and S^{ND} are :

$$
S_k^D = \frac{n_k}{h_1 n_k + 1} \left(\mu_0^D - \overline{x}_k^D\right) \left(\mu_0^D - \overline{x}_k^D\right)^T + \sum_{x_i^D \in C_k} \left(x_i^D - \overline{x}_k^D\right) \left(x_i^D - \overline{x}_k^D\right)^T,
$$

$$
S^{ND} = \frac{n}{h_0 n + 1} \left(\mu_0^{ND} - \overline{x}^{ND}\right) \left(\mu_0^{ND} - \overline{x}^{ND}\right)^T + \sum_{i=1}^{n} \left(x_i^{ND} - \overline{x}^{ND}\right) \left(x_i^{ND} - \overline{x}^{ND}\right)^T.
$$

On Heterogeneity, calculations are as follows :

$$
\begin{aligned}
f\left(X, y|G, w, \gamma\right) = \int & f\left(X, y|\gamma, G, w, \mu^{D}, \Sigma^{SD}, \mu^{ND}, \Sigma^{ND}\right) f\left(\mu^{D}|G, \gamma, \Sigma^{D}\right) f\left(\Sigma^{D}|G, \gamma\right) \\
& \times f\left(\mu^{ND}|\Sigma^{ND}, \gamma\right) f\left(\Sigma^{ND}|\gamma\right) d\mu^{D}\, d\Sigma^{D}\, d\mu^{ND}\, d\Sigma^{ND} \\
& = \int \prod_{k=1}^{G} \left[w_k^{n_k} \left(2\pi\right)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} \left|\Sigma_k^D\right|^{-\frac{(n_k+1)}{2}} \right] \\
& \times \exp\left[-\frac{1}{2} \sum_{k=1}^{G} \sum_{x_i \in k} \left(x_i^D - \mu_k^D\right)' \Sigma_k^{D^{-1}} \left(x_i^D - \mu_k^D\right)\right] \\
& \times \exp\left[-\frac{1}{2} \sum_{k=1}^{G} \left(\mu_k^D - \mu_0^D\right)' \left(h_1 \Sigma_k^D\right)^{-1} \left(\mu_k^D - \mu_0^D\right)\right] \\
& \times \prod_{k=1}^{G} \left\{ 2^{-\frac{p_\gamma(\delta+p_\gamma-1)}{2}} \pi^{-\frac{p_\gamma(p_\gamma-1)}{4}} \left|Q^D\right|^{\frac{(\delta+p_\gamma-1)}{2}} \left|\Sigma_k^D\right|^{-\frac{(\delta+2p_\gamma)}{2}} \right. \\
& \times \left. \left[\prod_{j=1}^{p_\gamma} \Gamma\left(\frac{\delta + p_\gamma - j}{2}\right)\right]^{-1} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma_k^{D^{-1}} Q^D\right)\right] \right\} \\
& \times \left(2\pi\right)^{-\frac{(n+1)(p-p_\gamma)}{2}} h_0^{-\frac{(p-p_\gamma)}{2}} \left|\Sigma^{ND}\right|^{-\frac{(n+1)}{2}} \\
& \times \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(x_i^{ND} - \mu^{ND}\right)' \Sigma^{ND^{-1}} \left(x_i^{ND} - \mu^{ND}\right)\right] \\
& \times \exp\left[-\frac{1}{2} \left(\mu^{ND} - \mu_0^{ND}\right)' \left(h_0 \Sigma^{ND}\right)^{-1} \left(\mu^{ND} - \mu_0^{ND}\right)\right] \\
& \times 2^{-\frac{(p-p_\gamma)(\delta+p-p_\gamma-1)}{2}} \pi^{-\frac{(p-p_\gamma)(p-p_\gamma-1)}{4}} \\
& \times \left[\prod_{j=1}^{p-p_\gamma} \Gamma\left(\frac{\delta + p - p_\gamma - j}{2}\right)\right]^{-1} \\
& \times \left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}} \left|\Sigma^{ND}\right|^{-\frac{[\delta+2(p-p_\gamma)]}{2}} \\
& \times \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{ND^{-1}} Q^{ND}\right)\right] d\mu^{D}\, d\Sigma^{D}\, d\mu^{ND}\, d\Sigma^{ND} \\
& = \int \prod_{k=1}^{G} \left\{ w_k^{n_k} \left(2\pi\right)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} \left|\Sigma_k^D\right|^{-\frac{(n_k+1)}{2}} \right. \\
& \times \left. \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma_k^{D^{-1}} S_k^D\right)\right] \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\times \exp\left\{ -\frac{1}{2}\sum_{k=1}^{G}\left(\mu_k^D - \frac{\sum_{x_i\in k} x_i^D + h_1^{-1}\mu_0^D}{n_k + h_1^{-1}}\right)' \left(n_k + h_1^{-1}\right)\Sigma_k^{D^{-1}}\right.\\
&\left.\times \left(\mu_k^D - \frac{\sum_{x_i\in k} x_i^D + h_1^{-1}\mu_0^D}{n_k + h_1^{-1}}\right)\right\}\\
&\times \prod_{k=1}^{G}\left\{2^{-\frac{p_\gamma(\delta+p_\gamma-1)}{2}}\pi^{-\frac{p_\gamma(p_\gamma-1)}{4}}\left|Q^D\right|^{\frac{(\delta+p_\gamma-1)}{2}}\left|\Sigma_k^D\right|^{-\frac{(\delta+2p_\gamma)}{2}}\right.\\
&\left.\times \left[\prod_{j=1}^{p_\gamma}\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\right]^{-1}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma_k^{D^{-1}}Q^D\right)\right]\right\}\\
&\times (2\pi)^{-\frac{(n+1)(p-p_\gamma)}{2}}h_0^{-\frac{(p-p_\gamma)}{2}}\left|\Sigma^{ND}\right|^{-\frac{(n+1)}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}}S^{ND}\right)\right]\\
&\times \exp\left\{ -\frac{1}{2}\left(\mu^{ND} - \frac{\sum_{i=1}^{n} x_i^{ND} + h_0^{-1}\mu_0^{ND}}{n + h_0^{-1}}\right)' \left(n + h_0^{-1}\right)\Sigma^{ND^{-1}}\right.\\
&\left.\times \left(\mu^{ND} - \frac{\sum_{i=1}^{n} x_i^{ND} + h_0^{-1}\mu_0^{ND}}{n + h_0^{-1}}\right)\right\}\\
&\times 2^{-\frac{(p-p_\gamma)(\delta+p-p_\gamma-1)}{2}}\pi^{-\frac{(p-p_\gamma)(p-p_\gamma-1)}{4}}\left[\prod_{j=1}^{p-p_\gamma}\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right]^{-1}\\
&\times \left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}}\left|\Sigma^{ND}\right|^{-\frac{[\delta+2(p-p_\gamma)]}{2}}\\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}}Q^{ND}\right)\right]d\mu^D\,d\Sigma^D\,d\mu^{ND}\,d\Sigma^{ND}\\
=\;&(2\pi)^{-\frac{np}{2}}\prod_{k=1}^{G}\left\{w_k^{n_k}\left(h_1 n_k + 1\right)^{-\frac{p_\gamma}{2}}\left|Q^D\right|^{\frac{(\delta+p_\gamma-1)}{2}}\right.\\
&\left.\times \left[\prod_{j=1}^{p_\gamma}\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\right]^{-1}\right\}\\
&\times 2^{-\frac{p_\gamma(\delta+p_\gamma-1)}{2}}\pi^{-\frac{p_\gamma(p_\gamma-1)}{4}}2^{-\frac{(p-p_\gamma)(\delta+p-p_\gamma-1)}{2}}\pi^{-\frac{(p-p_\gamma)(p-p_\gamma-1)}{4}}\\
&\times \left(h_0 n + 1\right)^{-\frac{(p-p_\gamma)}{2}}\left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}}\left[\prod_{j=1}^{p-p_\gamma}\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right]^{-1}\\
&\times \int\prod_{k=1}^{G}\left\{\left|\Sigma_k^D\right|^{\frac{-(n_k+\delta+2p_\gamma)}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma_k^{D^{-1}}\left[Q^D + S_k^D\right]\right)\right]\right\}\\
&\times \left|\Sigma^{ND}\right|^{-\frac{[n+\delta+2(p-p_\gamma)]}{2}}\\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}}\left[Q^{ND} + S^{ND}\right]\right)\right]d\Sigma^D\,d\Sigma^{ND}
\end{aligned}
$$

$$
\begin{aligned}
=&\pi^{\frac{-np}{2}}\prod_{k=1}^{G}\Bigg\{\left|Q^{D}\right|^{\frac{(\delta+p_\gamma-1)}{2}}\left|Q^{D}+S_k^{D}\right|^{\frac{-(n_k+\delta+p_\gamma-1)}{2}}\\
&\times w_k^{n_k}\left(h_1n_k+1\right)^{-\frac{p_\gamma}{2}}\prod_{j=1}^{p_\gamma}\frac{\Gamma\left(\frac{1}{2}\left(n_k+\delta+p_\gamma-j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta+p_\gamma-j\right)\right)}\Bigg\}\\
&\times\left|Q^{ND}\right|^{\frac{(\delta+p-p_\gamma-1)}{2}}\left|Q^{ND}+S^{ND}\right|^{\frac{-(n+\delta+p-p_\gamma-1)}{2}}\\
&\times\left(h_0n+1\right)^{-\frac{(p-p_\gamma)}{2}}\prod_{j=1}^{p-p_\gamma}\frac{\Gamma\left(\frac{1}{2}\left(n+\delta+p-p_\gamma-j\right)\right)}{\Gamma\left(\frac{1}{2}\left(\delta+p-p_\gamma-j\right)\right)}.
\end{aligned}
\qquad (A.2)
$$

Histograms, marginal posterior probabilities and clustering maps for the cases of ($\alpha_w = 1, \alpha_u = 6$), ($\alpha_w = 6, \alpha_u = 1$) and ($\alpha_w = 6, a_u = 6$), for the application of the simulated data set under section 2.5.

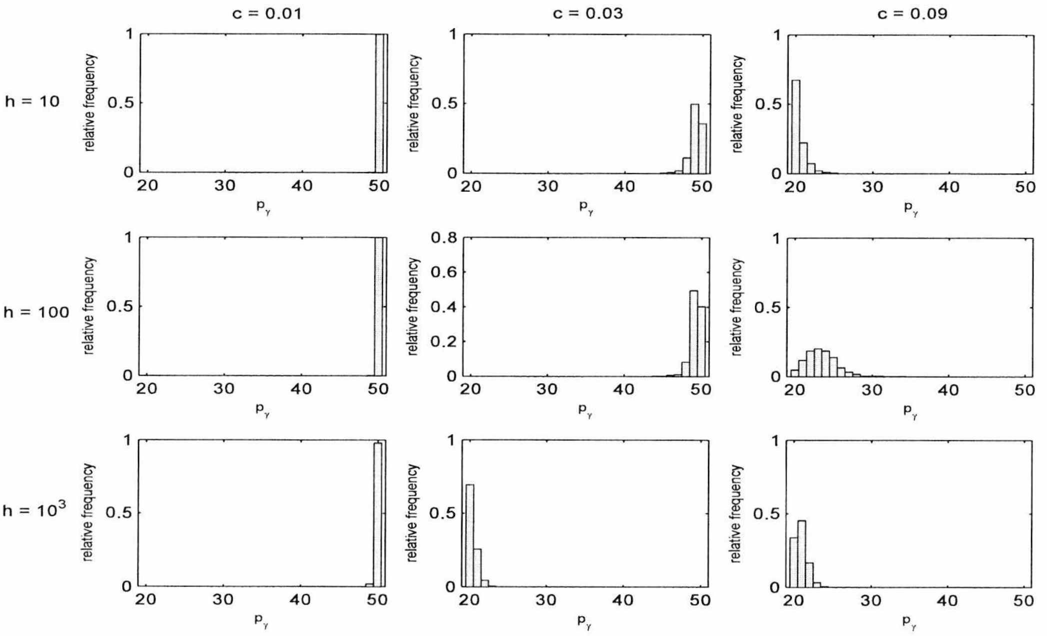

Figure A.0.1: Simulated data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 1$ and $a_u = 6$.

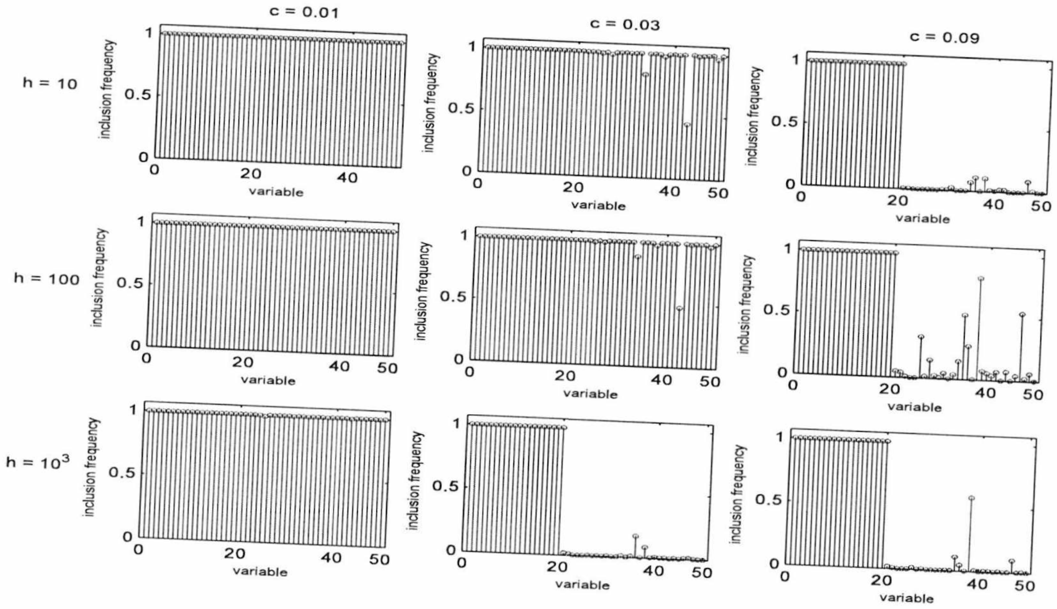

Figure A.0.2: Simulated data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 1$ and $a_u = 6$.

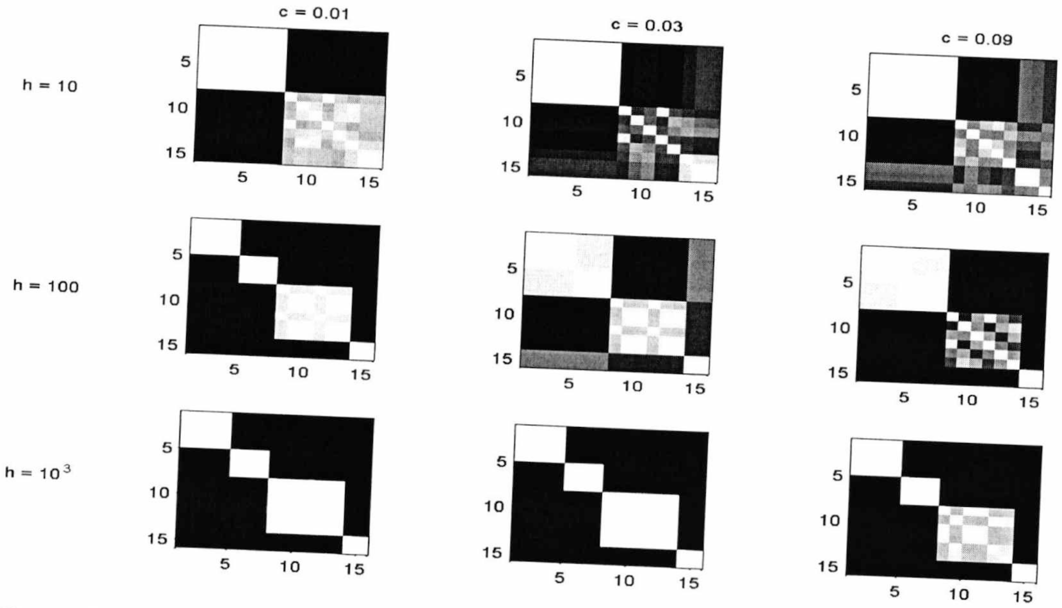

Figure A.0.3: Simulated data: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 1$ and $a_u = 6$.

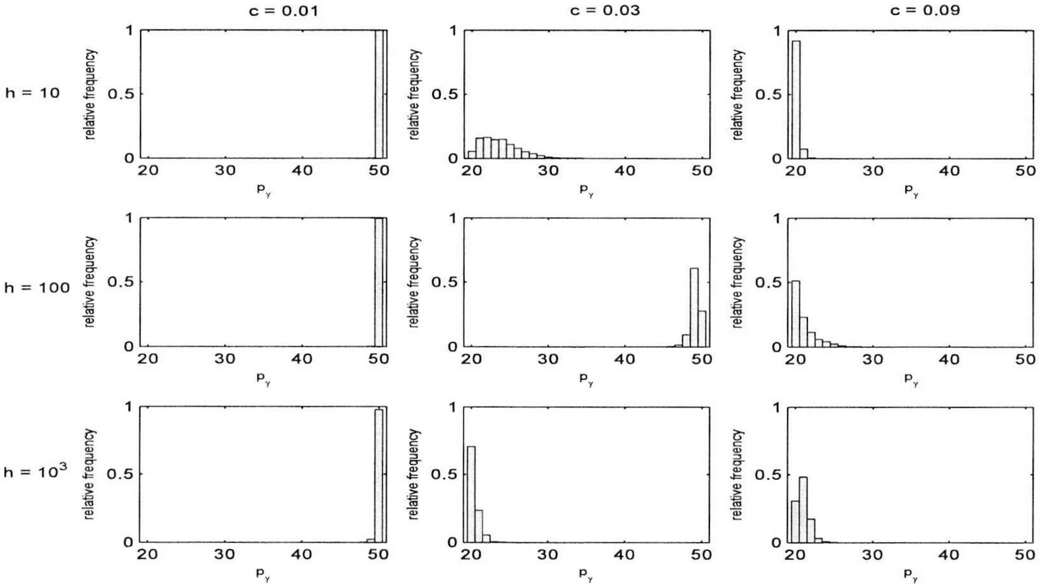

Figure A.0.4: Simulated data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 6$ and $a_u = 2$.

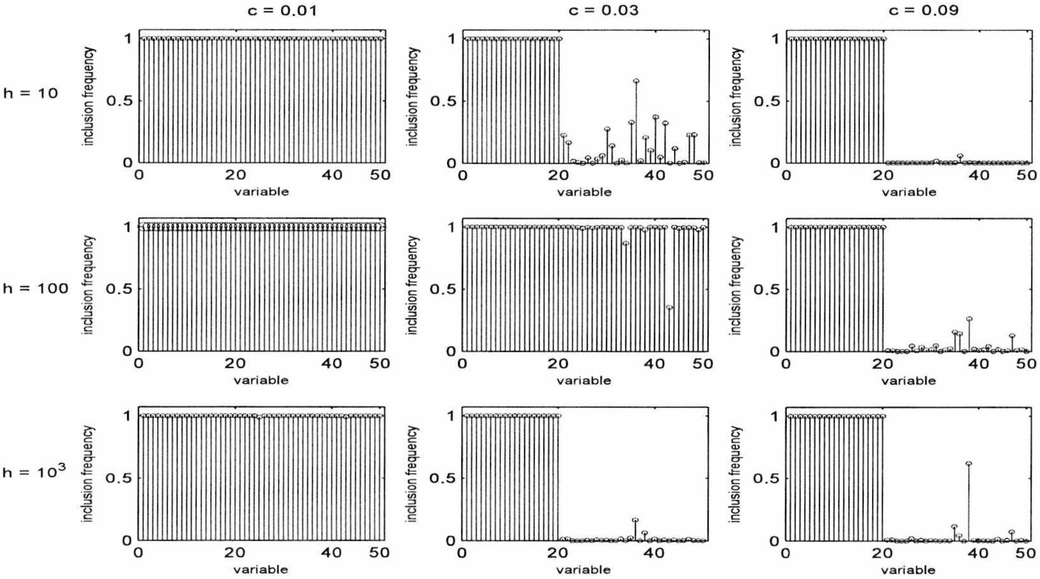

Figure A.0.5: Simulated data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 6$ and $a_u = 2$.

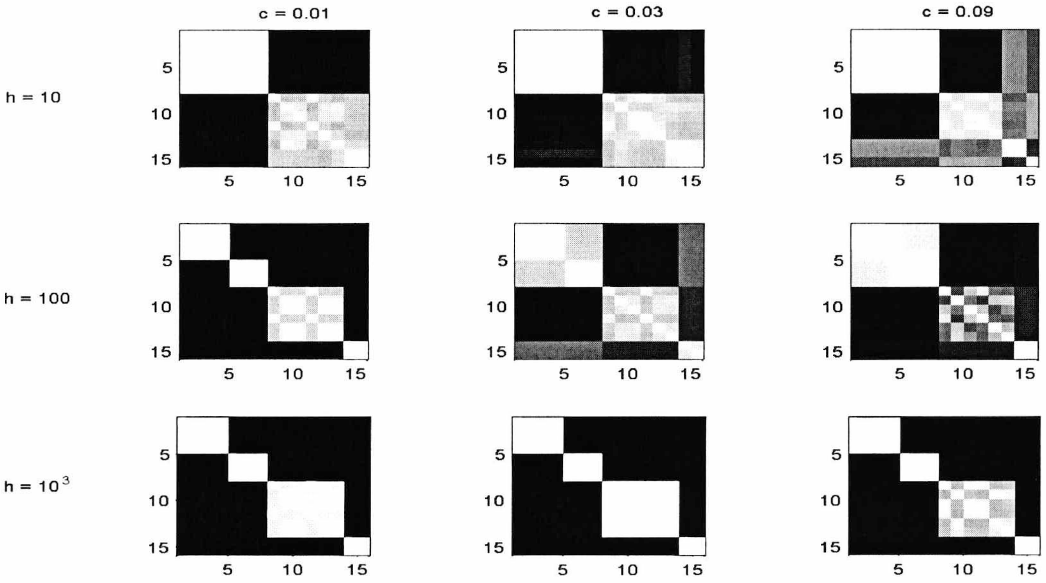

Figure A.0.6: Simulated data: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 6$ and $a_u = 2$.

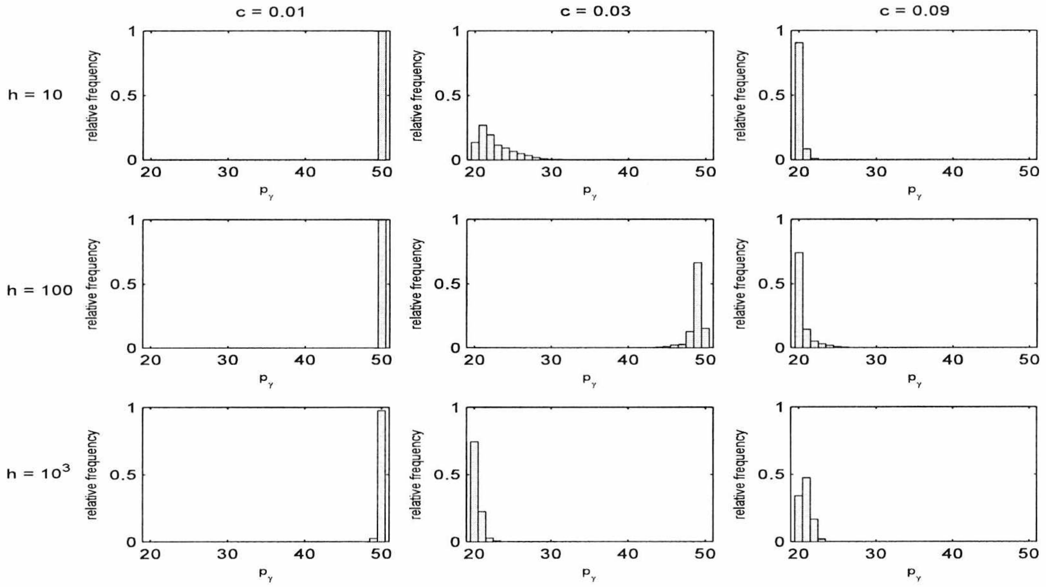

Figure A.0.7: Simulated data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 6$ and $a_u = 6$.

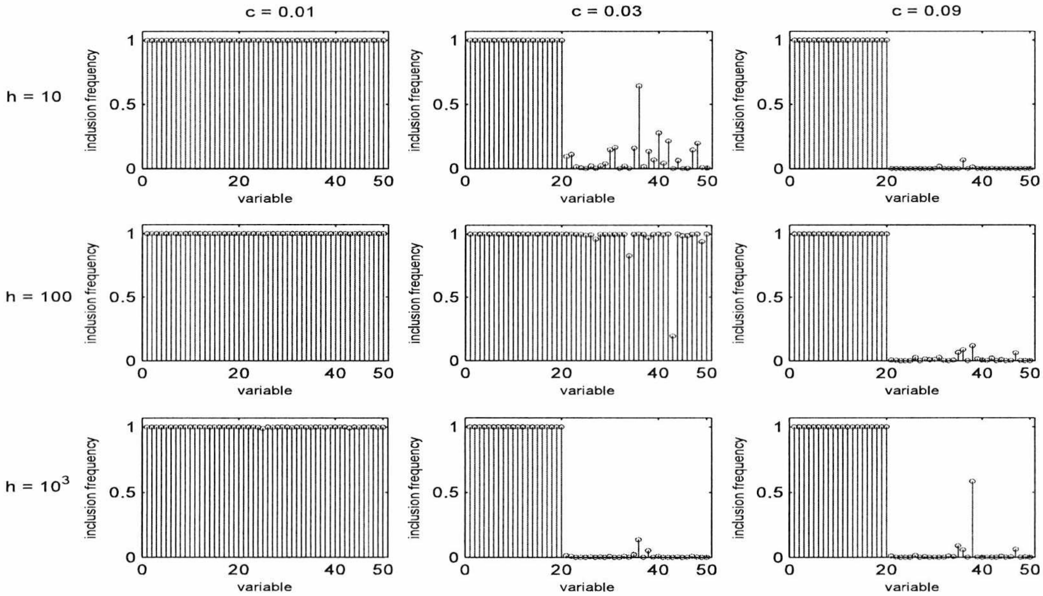

Figure A.0.8: Simulated data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 6$ and $a_u = 6$.

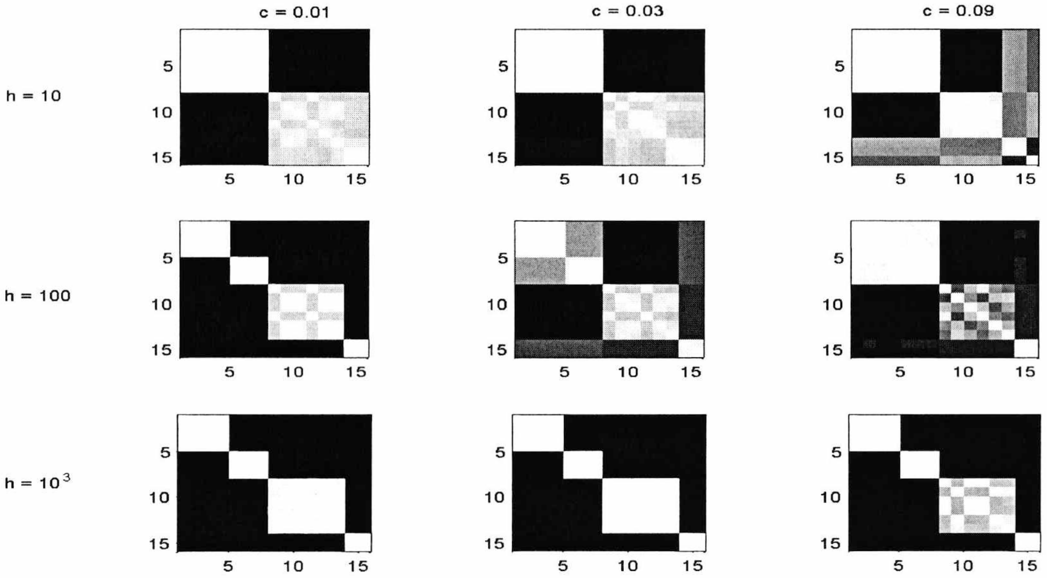

Figure A.0.9: Simulated data: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$, $c = 0.01, 0.03, 0.09$ and assuming unequal covariance matrices, when $\alpha_w = 6$ and $a_u = 6$.

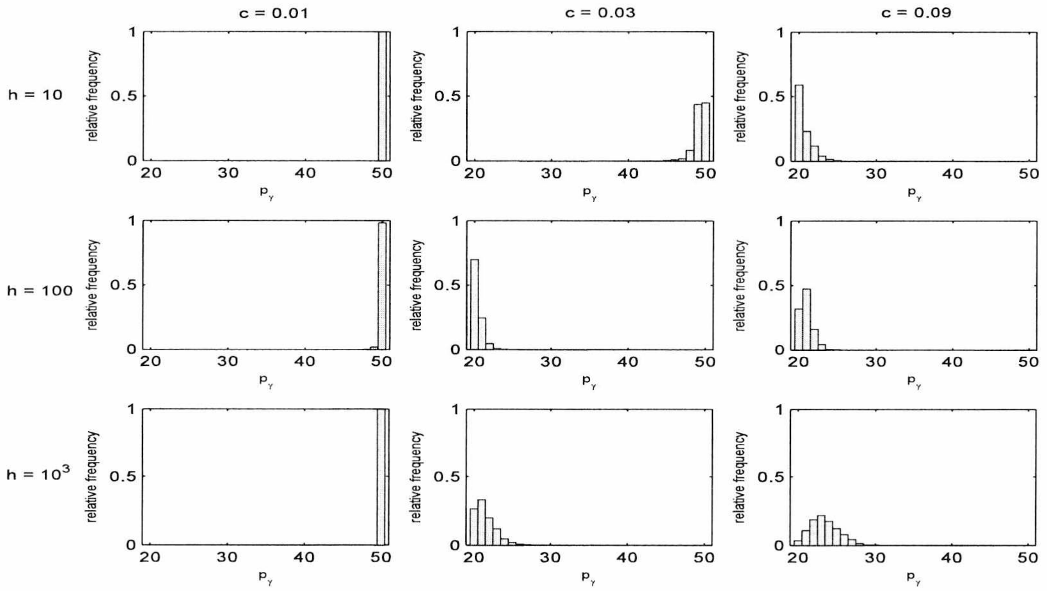

Figure A.0.10: Simulated data: Histograms of the total number of discriminating variables, p_γ, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and setting starting points the correct set of discriminating variables and cluster allocations, when $\alpha_w = 1$ and $\alpha_u = 2$.

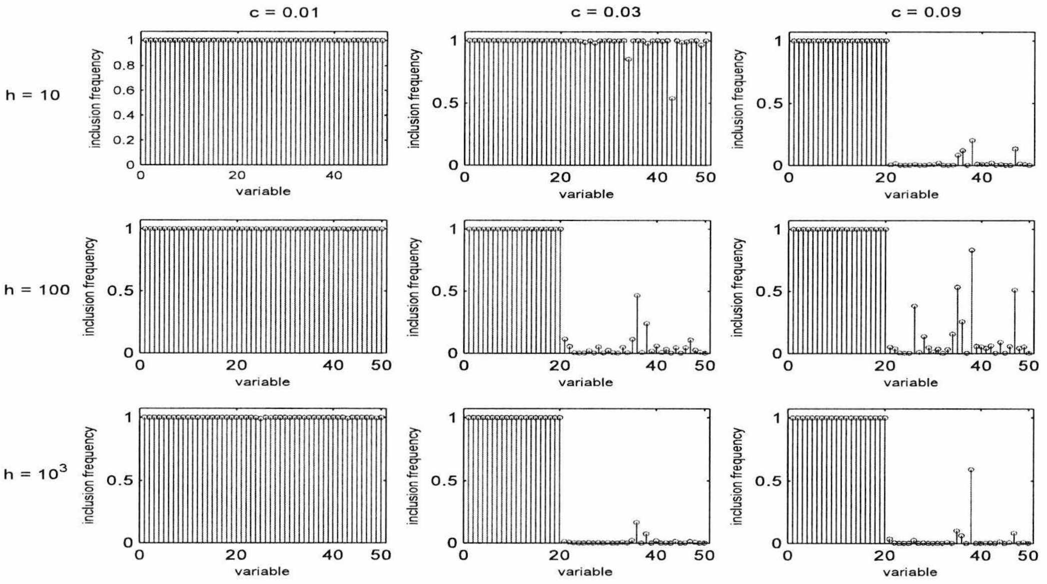

Figure A.0.11: Simulated data: Marginal Posterior Probabilities of the variables included in the model, for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and setting starting points the correct set of discriminating variables and cluster allocations, when $\alpha_w = 1$ and $\alpha_u = 2$.

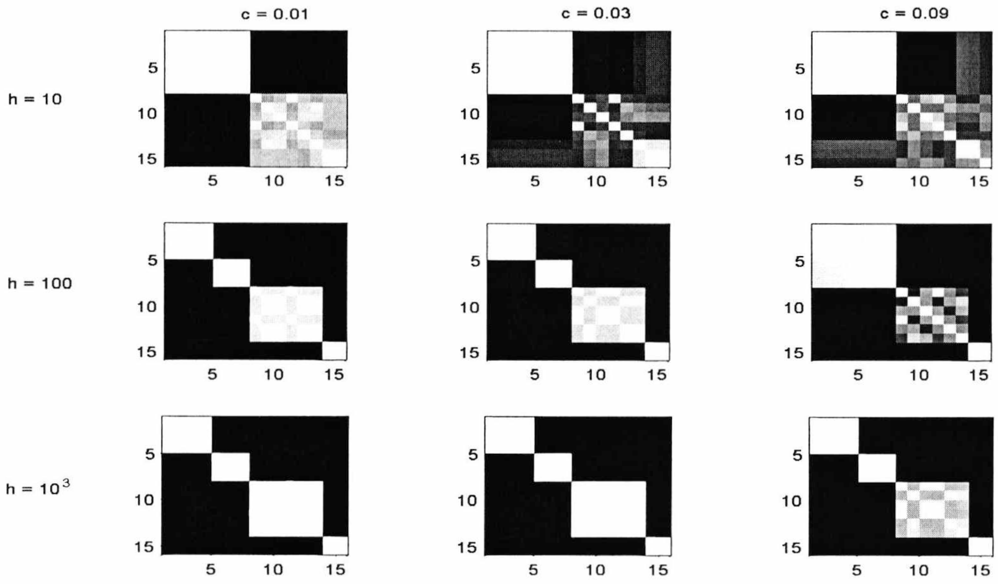

Figure A.0.12: Simulated data: Maps of the cluster allocations of the $n = 15$ observations for $h = 10, 100, 1000$ and $c = 0.01, 0.03, 0.09$, assuming unequal covariance matrices and setting starting points the correct set of discriminating variables and cluster allocations, when $\alpha_w = 1$ and $\alpha_u = 2$.

Dahl - Split Move

For the probability of observation t being allocated in component S^i we write :

$$Pr(y_t = S^i | y^{S^i}, y^{S^j}) = \frac{A}{A+B} \quad , \tag{A.3}$$

where the terms A, B are :

$$A = w_{S^i} \left(1 + \frac{h_1}{h_1 n_{S^i} + 1}\right)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \left[\frac{\Gamma\left(\frac{1}{2}(n_{S^i} + 1 + \delta + p_\gamma - j)\right)}{\Gamma\left(\frac{1}{2}(n_{S^i} + \delta + p_\gamma - j)\right)} \right]$$
$$\times \frac{\left|Q^D + S^D_{S^i}\right|_{\left(y^{S^i}, y_t\right)}^{-(n_{S^i}+1+\delta+p_\gamma-1)/2}}{\left|Q^D + S^D_{S^i}\right|_{\left(y^{S^i}\right)}^{-(n_{S^i}+\delta+p_\gamma-1)/2}},$$

$$B = w_{S^j}\left(1+\frac{h_1}{h_1 n_{S^j}+1}\right)^{-p_\gamma/2}\prod_{j=1}^{p_\gamma}\left[\frac{\Gamma\left(\frac{1}{2}(n_{S^j}+1+\delta+p_\gamma-j)\right)}{\Gamma\left(\frac{1}{2}(n_{S^j}+\delta+p_\gamma-j)\right)}\right]$$
$$\times\frac{\left|Q^D+S^D_{S^j}\right|^{-(n_{S^j}+1+\delta+p_\gamma-1)/2}_{(y^{S^j},y_t)}}{\left|Q^D+S^D_{S^j}\right|^{-(n_{S^j}+\delta+p_\gamma-1)/2}_{(y^{S^j})}}.$$

For a split move with acceptance probability min $(1, A)$, ratios A for the case of homogeneous and heterogeneous covariances are as follows :

1. Homogeneity.

$$A = \left[\frac{(h_1 n_{S^i}+1)(h_1 n_{S^j}+1)}{(h_1 n_S+1)}\right]^{\frac{-p_\gamma}{2}}\left[\frac{\left|Q^D+\sum_{k=1}^{G+1}S^D_k\right|}{\left|Q^D+\sum_{k=1}^{G}S^D_k\right|}\right]^{\frac{-(n+\delta+p_\gamma-1)}{2}}$$
$$\times\frac{1}{\prod_{t\in S^i}Pr(y_t=S^i|y^{S^i},y^{S^j})\prod_{t\in S^j}Pr(y_t=S^t|y^{S^i},y^{S^j})}$$
$$\times\frac{B(a_u,a_u)}{B(\alpha_w,G\alpha_w)}u^{n_{S^i}+\alpha_w-a_u}(1-u)^{n_{S^j}+\alpha_w-a_u}w_S^{\alpha_w} \qquad \text{(A.4)}$$

2. Heterogeneity.

$$A = \left|Q^D\right|^{\frac{\delta+p_\gamma-1}{2}}\frac{\left|QD+S^D_{S^i}\right|^{\frac{-(n_{S^i}+\delta+p_\gamma-1)}{2}}\left|Q^D+S^D_{S^j}\right|^{\frac{-(n_{S^j}+\delta+p_\gamma-1)}{2}}}{\left|Q^D+S^D_S\right|^{\frac{-(n_S+\delta+p_\gamma-1)}{2}}}$$
$$\times\left[\frac{(h_1 n_{S^i}+1)(h_1 n_{S^j}+1)}{(h_1 n_S+1)}\right]^{\frac{-p_\gamma}{2}}$$
$$\times\prod_{j=1}^{p_\gamma}\frac{\Gamma(\frac{1}{2}(n_{S^i}+\delta+p_\gamma-j))\Gamma(\frac{1}{2}(n_{S^j}+\delta+p_\gamma-j))}{\Gamma(\frac{1}{2}(n_S+\delta+p_\gamma-j))\Gamma(\frac{1}{2}(\delta+p_\gamma-j))}$$
$$\times\frac{1}{\prod_{t\in S^i}Pr(y_t=S^i|y^{S^i},y^{S^j})\prod_{t\in S^j}Pr(y_t=S^t|y^{S^i},y^{S^j})}$$
$$\times\frac{B(a_u,a_u)}{B(\alpha_w,G\alpha_w)}u^{n_{S^i}+\alpha_w-a_u}(1-u)^{n_{S^j}+\alpha_w-a_u}w_S^{\alpha_w} \qquad \text{(A.5)}$$

Appendix B

Appendix for Chapter 3

Calculations on the joint posterior $f\left(X, y|G, w, \gamma\right)$ for the Cluster Heterogeneity model considering Homogeneous Covariances :

$$\begin{aligned} f\left(X, y|G, w, \gamma\right) = \int f\left(X, y|\gamma, G, w, \mu^D, \Sigma^{SD}, \mu^{ND}, \Sigma^{ND}\right) f\left(\mu^D|G, \gamma, \Sigma^D\right) f\left(\Sigma^D|G, \gamma\right) \\ \times f\left(\mu^{ND}|\Sigma^{ND}, \gamma\right) f\left(\Sigma^{ND}|\gamma\right) d\mu^D \, d\Sigma^D \, d\mu^{ND} \, d\Sigma^{ND}. \end{aligned} \tag{B.1}$$

However, the covariance matrices are now multiples of the Identity matrix, i.e. $\Sigma^D = cI_{p_\gamma}$ and $\Sigma^{ND} = dI_{p-p_\gamma}$. Therefore, integration of the posterior over the covariance matrices simplifies into integrating over parameters c, d. Hence, we write:

$$\begin{aligned} f\left(X, y|G, w, \gamma\right) = \int \prod_{k=1}^{G} \left[w_k^{n_k} \left(2\pi\right)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} \left|\Sigma^D\right|^{-\frac{(n_k+1)}{2}} \right] \\ \times \exp\left[-\frac{1}{2} \sum_{k=1}^{G} \sum_{x_i \in k} \left(x_i^D - \mu_k^D\right)' \Sigma^{D^{-1}} \left(x_i^D - \mu_k^D\right) \right] \\ \times \exp\left[-\frac{1}{2} \sum_{k=1}^{G} \left(\mu_k^D - \mu_0^D\right)' \left(h_1 \Sigma^D\right)^{-1} \left(\mu_k^D - \mu_0^D\right) \right] \\ \times \frac{{\beta_c}^{\alpha_c}}{\Gamma\left(\alpha_c\right)} c^{-(\alpha_c+1)} \exp\left(-\beta_c c^{-1}\right) \end{aligned}$$

$$
\begin{aligned}
&\times (2\pi)^{-\frac{(n+1)(p-p_\gamma)}{2}} h_0^{-\frac{(p-p_\gamma)}{2}} \left|\Sigma^{ND}\right|^{-\frac{(n+1)}{2}} \\
&\times \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(x_i^{ND}-\mu^{ND}\right)'\Sigma^{ND-1}\left(x_i^{ND}-\mu^{ND}\right)\right] \\
&\times \exp\left[-\frac{1}{2}\left(\mu^{ND}-\mu_0^{ND}\right)'\left(h_0\Sigma^{ND}\right)^{-1}\left(\mu^{ND}-\mu_0^{ND}\right)\right] \\
&\times \frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} d^{-(\alpha_d+1)} \exp\left(-\beta_d d^{-1}\right) d\mu^D\, d\mu^{ND}\, dc\, dd \\
=& \int \prod_{k=1}^{G} \Bigg\{ w_k^{n_k} (2\pi)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} c^{-\frac{p_\gamma(n_k+1)}{2}} \\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{D-1}S_k^D\right)\right] \Bigg\} \\
&\times \exp\Bigg\{-\frac{1}{2}\sum_{k=1}^{G}\left(\mu_k^D-\frac{\sum_{x_i\in k}x_i^D+h_1^{-1}\mu_0^D}{n_k+h_1^{-1}}\right)'\left(n_k+h_1^{-1}\right)\Sigma^{D-1} \\
&\times \left(\mu_k^D-\frac{\sum_{x_i\in k}x_i^D+h_1^{-1}\mu_0^D}{n_k+h_1^{-1}}\right)\Bigg\} \\
&\times \frac{{\beta_c}^{\alpha_c}}{\Gamma\left(\alpha_c\right)} c^{-(\alpha_c+1)} \exp\left(-\beta_c c^{-1}\right) \\
&\times (2\pi)^{-\frac{(n+1)(p-p_\gamma)}{2}} h_0^{-\frac{(p-p_\gamma)}{2}} d^{-\frac{(p-p_\gamma)(n+1)}{2}} \\
&\times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND-1}S^{ND}\right)\right] \\
&\times \exp\Bigg\{-\frac{1}{2}\left(\mu^{ND}-\frac{\sum_{i=1}^{n}x_i^{ND}+h_0^{-1}\mu_0^{ND}}{n+h_0^{-1}}\right)'\left(n+h_0^{-1}\right)\Sigma^{ND-1} \\
&\times \left(\mu^{ND}-\frac{\sum_{i=1}^{n}x_i^{ND}+h_0^{-1}\mu_0^{ND}}{n+h_0^{-1}}\right)\Bigg\} \\
&\times \frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} d^{-(\alpha_d+1)} \exp\left(-\beta_d d^{-1}\right) d\mu^D\, d\mu^{ND}\, dc\, dd \\
=& (2\pi)^{-\frac{np}{2}} \left(h_0 n+1\right)^{-\frac{(p-p_\gamma)}{2}} \prod_{k=1}^{G}\left[w_k^{n_k}\left(h_1 n_k+1\right)^{-\frac{p_\gamma}{2}}\right] \\
&\times \frac{{\beta_c}^{\alpha_c}}{\Gamma\left(\alpha_c\right)}\frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} \\
&\times \int c^{-\left(\alpha_c+1+\frac{np_\gamma}{2}\right)} \exp\Bigg\{-c^{-1}\left[\beta_c+\frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{G}S_k^D\right)\right]\Bigg\} \\
&\times d^{-\left(\alpha_d+1+\frac{n(p-p_\gamma)}{2}\right)} \exp\Bigg\{-d^{-1}\left[\beta_d+\frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{G}S^{ND}\right)\right]\Bigg\} dc\, dd
\end{aligned}
$$

$$
\begin{aligned}
&= (2\pi)^{-\frac{np}{2}} \left(h_0 n+1\right)^{-\frac{(p-p_\gamma)}{2}} \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k+1\right)^{-\frac{p_\gamma}{2}}\right] \\
&\times \frac{{\beta_c}^{\alpha_c}}{\Gamma\left(\alpha_c\right)} \Gamma\left(\alpha_c+\frac{np_\gamma}{2}\right) \left[\beta_c+\frac{1}{2}\operatorname{tr}\left(\sum_{k=1}^{G} S_k^D\right)\right]^{-\left(\alpha_c+\frac{np_\gamma}{2}\right)} \\
&\times \frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} \Gamma\left(\alpha_d+\frac{n\left(p_\gamma\right)}{2}\right) \left[\beta_d+\frac{1}{2}\operatorname{tr}\left(S^{ND}\right)\right]^{-\left(\alpha_d+\frac{n(p-p_\gamma)}{2}\right)},
\end{aligned}
\tag{B.2}
$$

where S_k^D and S^{ND} are :

$$
S_k^D = \frac{n_k}{h_1 n_k+1}\left(\mu_0^D-\overline{x}_k^D\right)\left(\mu_0^D-\overline{x}_k^D\right)^T + \sum_{x_i^D \in C_k}\left(x_i^D-\overline{x}_k^D\right)\left(x_i^D-\overline{x}_k^D\right)^T,
$$

$$
S^{ND} = \frac{n}{h_0 n+1}\left(\mu_0^{ND}-\overline{x}^{ND}\right)\left(\mu_0^{ND}-\overline{x}^{ND}\right)^T + \sum_{i=1}^{n}\left(x_i^{ND}-\overline{x}^{ND}\right)\left(x_i^{ND}-\overline{x}^{ND}\right)^T.
$$

On Heterogeneity, with $\Sigma_k^D = c_k I_{p_\gamma}$ and $\Sigma^{ND} = dI_{p-p_\gamma}$, calculations are as follows:

$$
\begin{aligned}
f\left(X, y|G, w, \gamma\right) = \int \prod_{k=1}^{G} &\left[w_k^{n_k} (2\pi)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} \left|\Sigma_k^D\right|^{-\frac{(n_k+1)}{2}}\right] \\
&\times \exp\left[-\frac{1}{2}\sum_{k=1}^{G}\sum_{x_i \in k}\left(x_i^D-\mu_k^D\right)' {\Sigma_k^D}^{-1}\left(x_i^D-\mu_k^D\right)\right] \\
&\times \exp\left[-\frac{1}{2}\sum_{k=1}^{G}\left(\mu_k^D-\mu_0^D\right)'\left(h_1\Sigma_k^D\right)^{-1}\left(\mu_k^D-\mu_0^D\right)\right] \\
&\times \frac{{\beta_c}^{G\alpha_c}}{\Gamma\left(\alpha_c\right)^G}\prod_{k=1}^{G}\left[c_k^{-(\alpha_c+1)}\right]\exp\left[-\beta_c\sum_{k=1}^{G} c_k^{-1}\right] \\
&\times (2\pi)^{-\frac{(n+1)(p-p_\gamma)}{2}} h_0^{-\frac{(p-p_\gamma)}{2}} \left|\Sigma^{ND}\right|^{-\frac{(n+1)}{2}} \\
&\times \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(x_i^{ND}-\mu^{ND}\right)' {\Sigma^{ND}}^{-1}\left(x_i^{ND}-\mu^{ND}\right)\right] \\
&\times \exp\left[-\frac{1}{2}\left(\mu^{ND}-\mu_0^{ND}\right)'\left(h_0\Sigma^{ND}\right)^{-1}\left(\mu^{ND}-\mu_0^{ND}\right)\right] \\
&\times \frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} d^{-(\alpha_d+1)} \exp\left(-\beta_d d^{-1}\right) d\mu^D \, d\mu^{ND} \, dc_k \, dd
\end{aligned}
$$

$$
\begin{aligned}
&= \int \prod_{k=1}^{G} \left\{ w_k^{n_k} \left(2\pi\right)^{-\frac{(n_k+1)p_\gamma}{2}} h_1^{-\frac{p_\gamma}{2}} c_k^{-\frac{p_\gamma(n_k+1)}{2}} \right. \\
&\quad \left. \times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma_k^{D^{-1}} S_k^D\right)\right] \right\} \\
&\quad \times \exp\Bigg\{ -\frac{1}{2}\sum_{k=1}^{G} \left(\mu_k^D - \frac{\sum_{x_i\in k} x_i^D + h_1^{-1}\mu_0^D}{n_k + h_1^{-1}}\right)' \left(n_k + h_1^{-1}\right) \Sigma_k^{D^{-1}} \\
&\quad \times \left(\mu_k^D - \frac{\sum_{x_i\in k} x_i^D + h_1^{-1}\mu_0^D}{n_k + h_1^{-1}}\right) \Bigg\} \\
&\quad \times \frac{{\beta_c}^{G\alpha_c}}{\Gamma\left(\alpha_c\right)^G} \prod_{k=1}^{G} \left[c_k^{-(\alpha_c+1)}\right] \exp\left[-\beta_c \sum_{k=1}^{G} c_k^{-1}\right] \\
&\quad \times \left(2\pi\right)^{-\frac{(n+1)(p-p_\gamma)}{2}} h_0^{-\frac{(p-p_\gamma)}{2}} d^{-\frac{(p-p_\gamma)(n+1)}{2}} \\
&\quad \times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma^{ND^{-1}} S^{ND}\right)\right] \\
&\quad \times \exp\Bigg\{ -\frac{1}{2}\left(\mu^{ND} - \frac{\sum_{i=1}^{n} x_i^{ND} + h_0^{-1}\mu_0^{ND}}{n + h_0^{-1}}\right)' \left(n + h_0^{-1}\right) \Sigma^{ND^{-1}} \\
&\quad \times \left(\mu^{ND} - \frac{\sum_{i=1}^{n} x_i^{ND} + h_0^{-1}\mu_0^{ND}}{n + h_0^{-1}}\right) \Bigg\} d\mu^D \, d\mu^{ND} \, dc_k \, dd \\
&= \left(2\pi\right)^{-\frac{np}{2}} \left(h_0 n + 1\right)^{-\frac{(p-p_\gamma)}{2}} \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-\frac{p_\gamma}{2}}\right] \\
&\quad \times \frac{{\beta_c}^{\alpha_c}}{\Gamma\left(\alpha_c\right)} \frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} \\
&\quad \times \int \prod_{k}^{G} \left\{ c^{-\left(\alpha_c + 1 + \frac{n_k p_\gamma}{2}\right)} \exp\left[-c^{-1}\left(\beta_c + \frac{1}{2}\mathrm{tr}\left[\sum_{k=1}^{G} S_k^D\right]\right)\right] \right\} \\
&\quad \times d^{-\left(\alpha_d + 1 + \frac{n(p-p_\gamma)}{2}\right)} \exp\left[-d^{-1}\left(\beta_d + \frac{1}{2}\mathrm{tr}\left[\sum_{k=1}^{G} S^{ND}\right]\right)\right] dc_k \, dd \\
&= \left(2\pi\right)^{-\frac{np}{2}} \left(h_0 n + 1\right)^{-\frac{(p-p_\gamma)}{2}} \prod_{k=1}^{G} \left[w_k^{n_k} \left(h_1 n_k + 1\right)^{-\frac{p_\gamma}{2}}\right] \\
&\quad \times \frac{{\beta_c}^{G\alpha_c}}{\Gamma\left(\alpha_c\right)^G} \prod_{k=1}^{G} \left\{ \Gamma\left(\alpha_c + \frac{n_k p_\gamma}{2}\right) \left[\beta_c + \frac{1}{2}\mathrm{tr}\left(S_k^D\right)\right]^{-\left(\alpha_c + \frac{n_k p_\gamma}{2}\right)} \right\} \\
&\quad \times \frac{{\beta_d}^{\alpha_d}}{\Gamma\left(\alpha_d\right)} \Gamma\left(\alpha_d + \frac{n\left(p - p_\gamma\right)}{2}\right) \left[\beta_d + \frac{1}{2}\mathrm{tr}\left(S^{ND}\right)\right]^{-\left(\alpha_d + \frac{n(p-p_\gamma)}{2}\right)}
\end{aligned}
\tag{B.3}
$$

Appendix C

Appendix for Chapter 4

Trace plots, histograms, marginal posterior probabilities and clustering maps for the applications of the CH, CVH, CVH with a prior on the mean value m, CVH with priors on m and h and Non-Conjugate models on the iris, crabs and arthritis data sets.

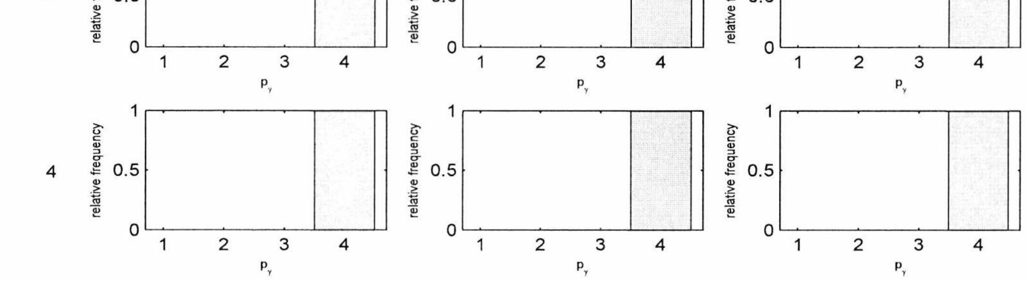

Figure C.0.13: Cluster-Variable Heterogeneity with a prior on m for the Iris data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 1000$.

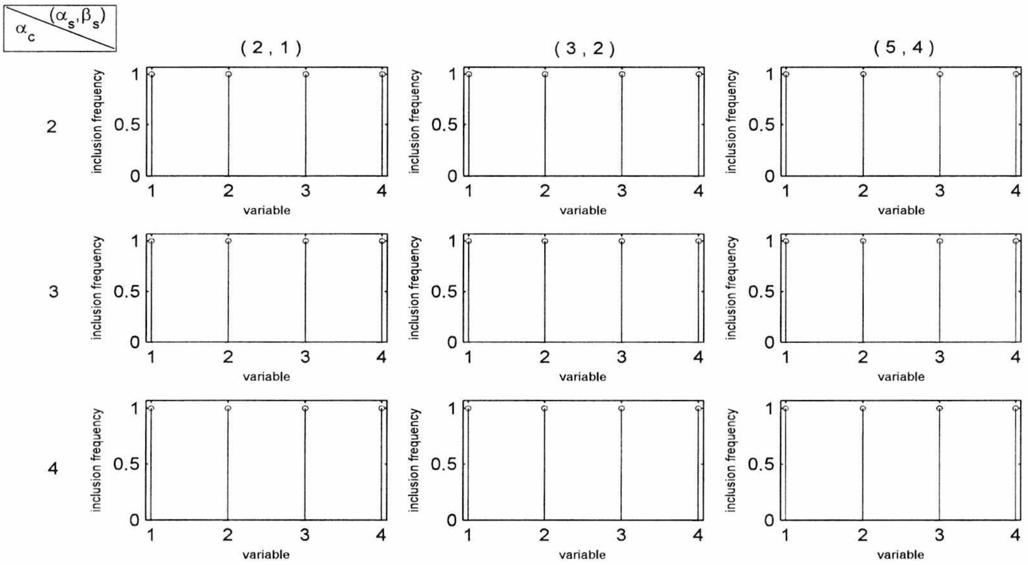

Figure C.0.14: Cluster-Variable Heterogeneity with a prior on m for the Iris data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, assuming heterogeneity and setting $h = 1000$.

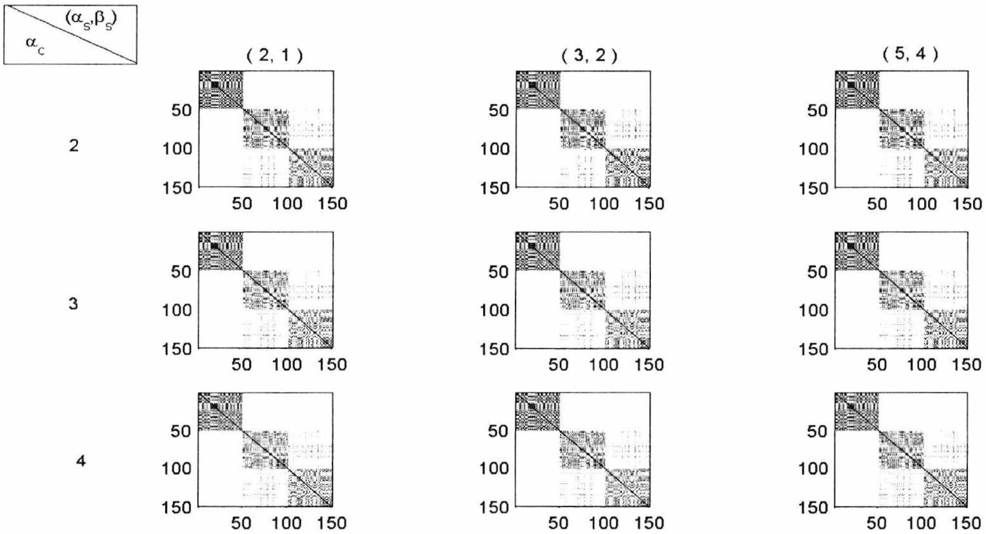

Figure C.0.15: Cluster-Variable Heterogeneity with a prior on m for the Iris data set: Maps of the cluster allocations of the $n = 150$ observations for various combinations of (α_s, β_s) and α_c assuming heterogeneity and setting $h = 1000$.

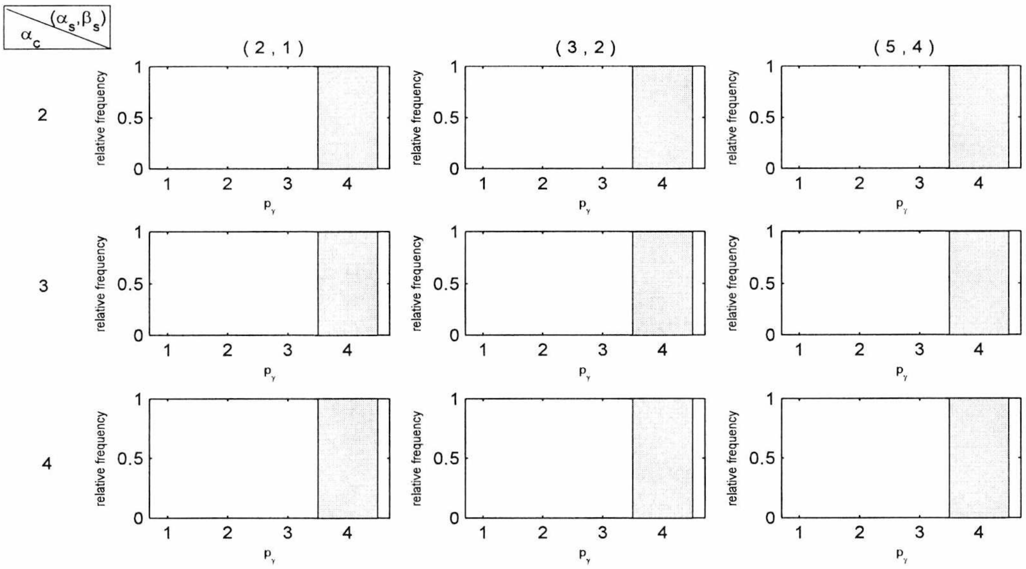

Figure C.0.16: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

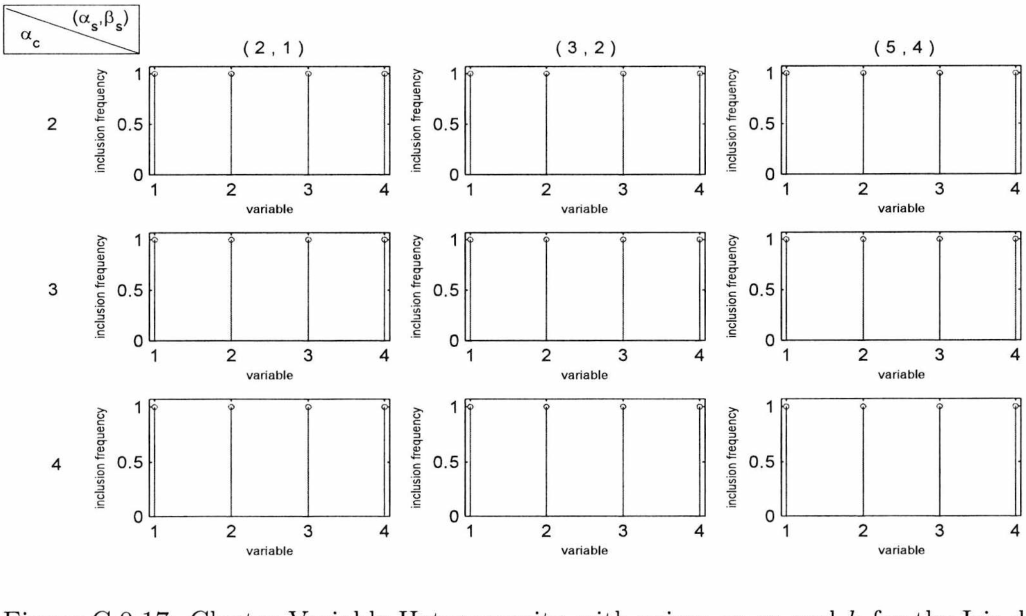

Figure C.0.17: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

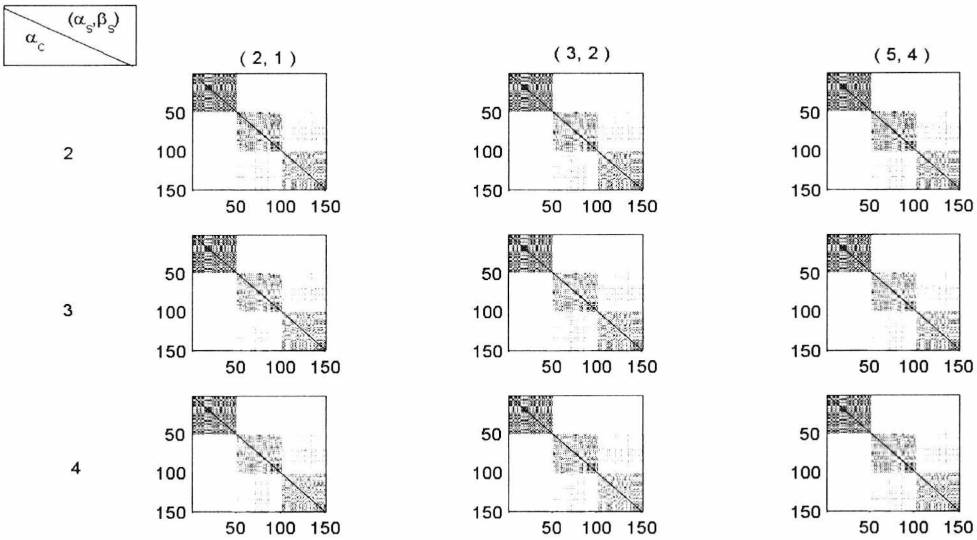

Figure C.0.18: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Maps of the cluster allocations of the $n = 150$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

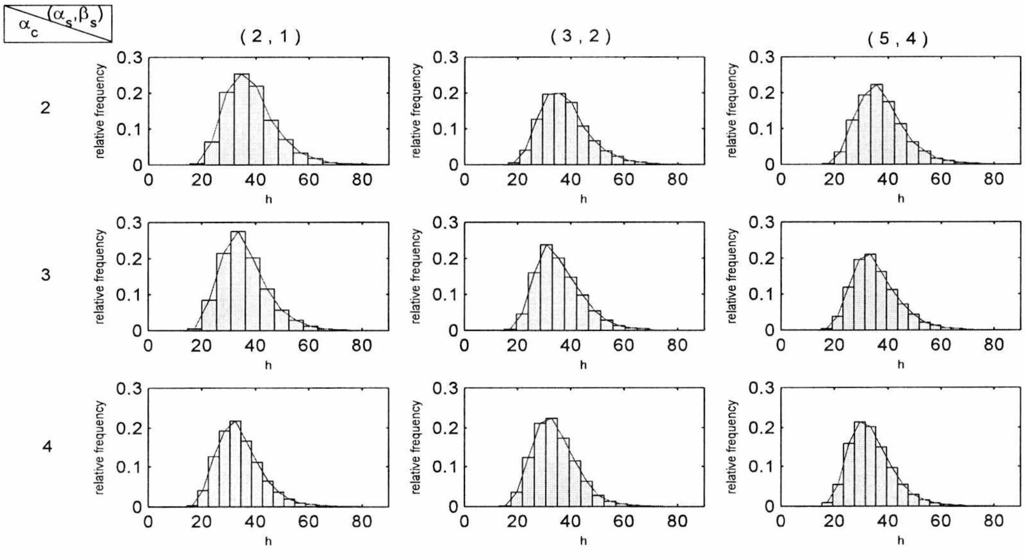

Figure C.0.19: Cluster-Variable Heterogeneity with priors on m and h for the Iris data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

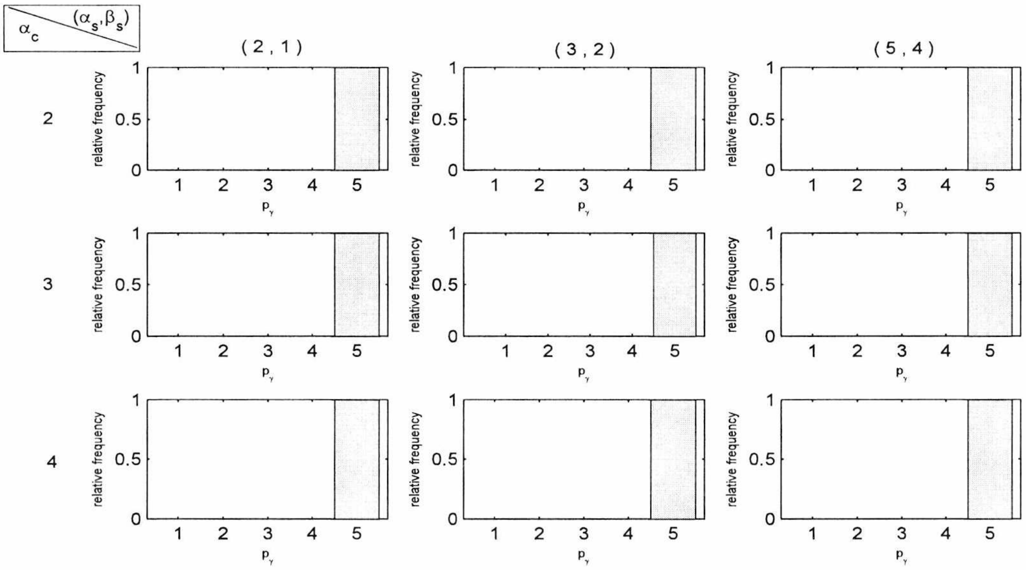

Figure C.0.20: Cluster-Variable Heterogeneity with priors on m and h for the principal component of the Crabs data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

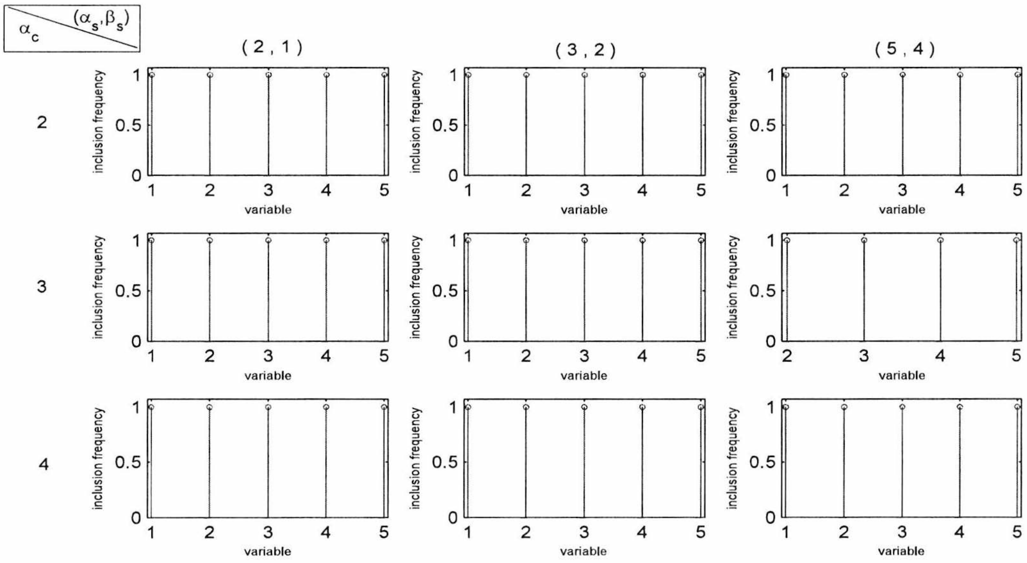

Figure C.0.21: Cluster-Variable Heterogeneity with priors on m and h for the principal components of the Crabs data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

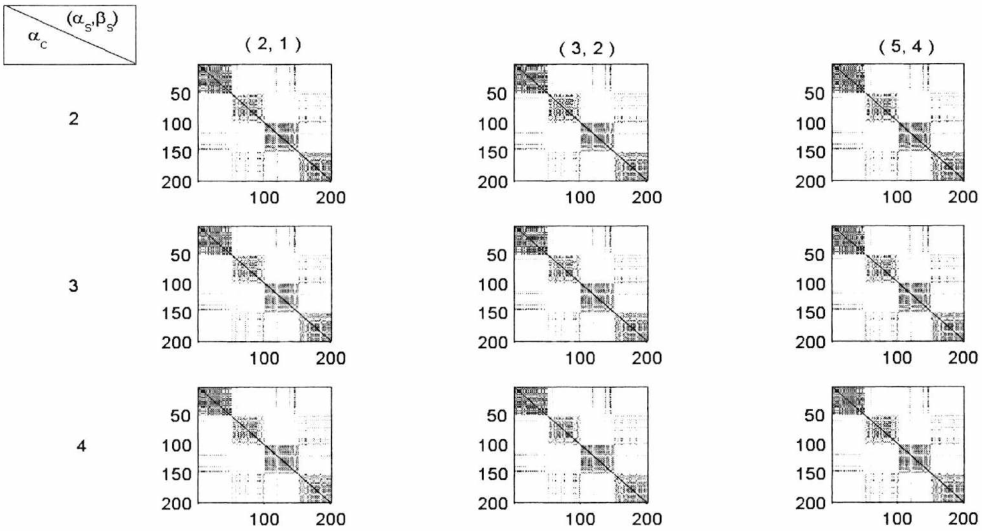

Figure C.0.22: Cluster-Variable Heterogeneity with priors on m and h for the principal of the Crabs data set: Maps of the cluster allocations of the $n = 200$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

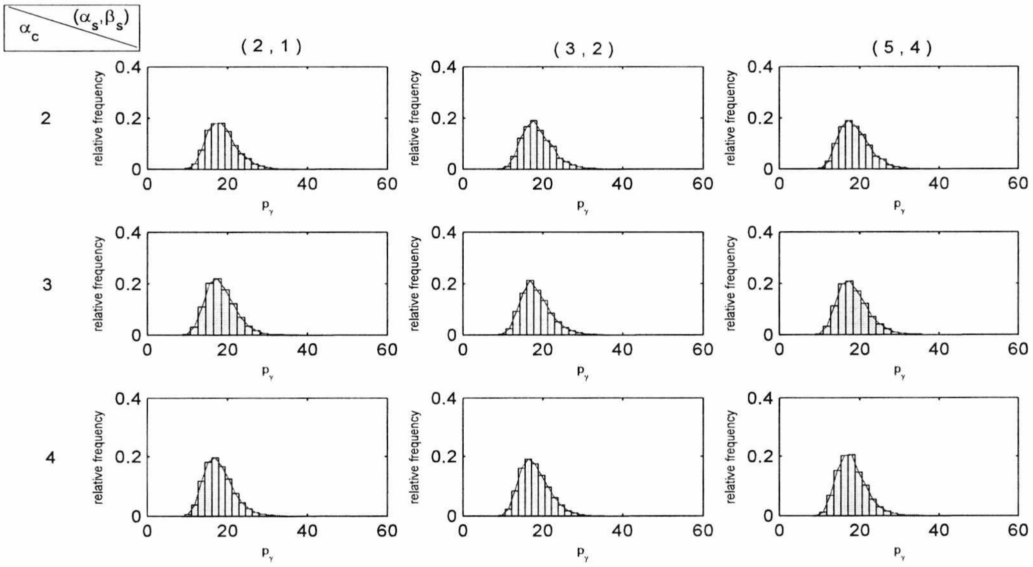

Figure C.0.23: Cluster-Variable Heterogeneity with priors on m and h for the principal components of the Crabs data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

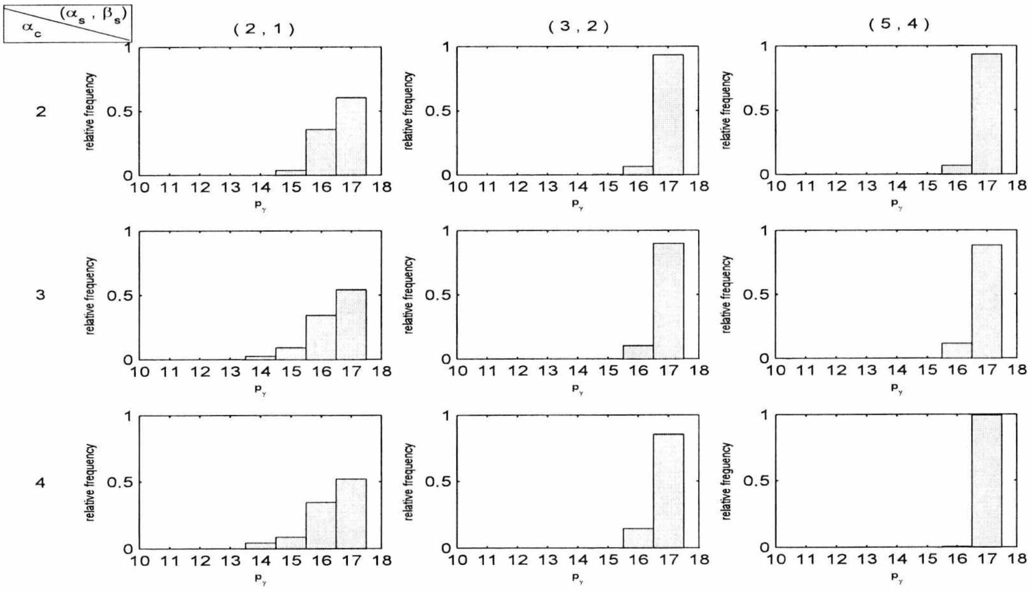

Figure C.0.24: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Histograms of the total number of discriminating variables, p_γ, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

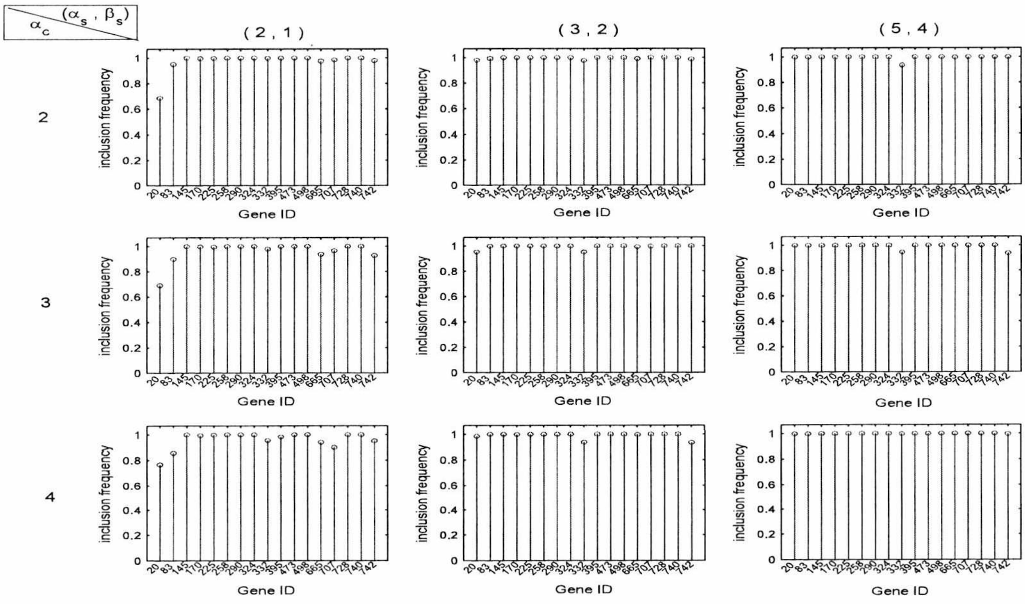

Figure C.0.25: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Marginal Posterior Probabilities of the variables included in the model, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

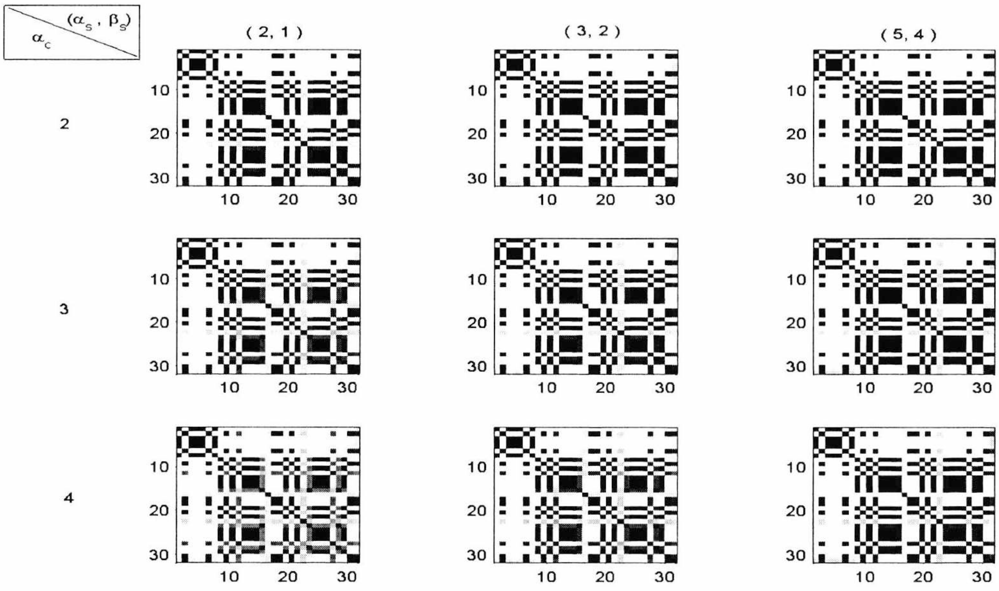

Figure C.0.26: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Maps of the cluster allocations of the $n = 31$ observations for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

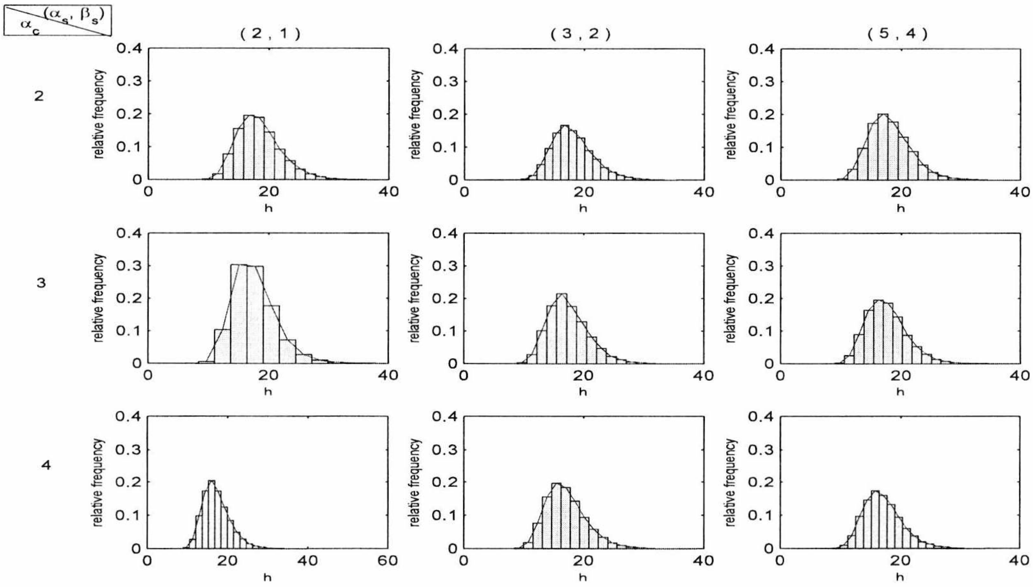

Figure C.0.27: Cluster-Variable Heterogeneity with priors on m and h for the set of 17 variables of the arthritis data set: Posterior distribution of h, for various combinations of (α_s, β_s) and α_c, under the case of $h \sim IG(3, 400)$, with unequal covariance matrices.

Bibliography

Anderson, E. (1935). The Irises of the Gaspe Peninsula. *Bulletin of the American Iris Society 59*, 2–5.

Anderson, E. (1936). The Species Problem in Iris. *Annals of the Missouri Botanical Garden 23*, 457–509.

Breyer, L. A. and G. O. Roberts (2000). From Metropolis to Diffusions: Gibbs States and Optimal Scaling. *Stochastic Processes and their Applications 90*, 181–206.

Brooks, S. P. and G. O. Roberts (1998). Convergence Assessment Techniques for Markov Chain Monte Carlo. *Statistics and Computing 8*, 319–335.

Brown, P. J. (1993). *Measurement, Regression, and Calibration.* Oxford, U.K.

Brown, P. J., M. Vannucci, and T. Fearn (1998). Multivariate Bayesian Variable Selection and Prediction. *Journal of the Royal Statistical Society Series B, 60*, 627–641.

Brusco, M. J. and J. D. Cradit (2001). A Variable Selection Heuristic for K-means Clustering. *Psychometrika 66*, 249–270.

Carmone JR, F. J., A. Kara, and S. Maxwell (1999). "HINoV" : A New Model to Improve Market Segment Definition by Identifying Noisy Variables. *Journal of Marketing Research 36*, 501–509.

Celeux, G. (1998). Bayesian Inference for Mixtures: The Label Switching Problem. In *COMPSTAT 98 (eds.R. Payne and P. Green)*, pp. 227–232. Heidelberg : Physica.

Celeux, G., M. Hurn, and C. P. Roberts (2000). Computational and Iinferential Difficulties With Mixture Posterior Distributions. *Journal of the American Statistical Association 95*, 957–970.

Cowles, M. K. and B. P. Carlin (1996). Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review. *Journal of the American Statistical Association 91*, 883–904.

Cramer, R. (1946). *Mathematical Methods of Statistics.* Princeton.

Dahl, D. B. (2005). Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models. Technical report, Texas A&M University.

Dahl, D. B. (2006). Modelling Differential Gene Expression Using a Dirichlet Process Mixture Model. In *Bayesian Inference for Gene Expression and Proteomics.* Do, K.A., and Müller, P., and Vannucci, M.

Dellaportas, P., J. J. Forster, and I. Ntzoufras (2000). *Bayesian Variable Selection Using the Gibbs Sampler*, Chapter Generalized Linear Models : A Bayesian Perspective., pp. 273–286. New York, USA, Chemical Rubber Company Press.

Diebolt, J. and C. P. Robert (1994). Estimation of Finite Mixture Distributions Through Bayesian Sampling. *Journal of the Royal Statistical Society Series B, 56*, 363–375.

Everitt, B. S., S. Landau, and M. Leese (2001). *Cluster Analysis.* Arnold.

Fowlkes, E. B., R. Gnanadesikan, and J. R. Kettering (1988). Variable Selection in Clustering. *Journal of Classification 5*, 205–228.

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

Friedman, J. H. and J. J. Meulman (2004). Clustering Objects on Subsets of Variables. *Journal of the Royal Statistical Society Series B*, 1–25.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association 85*, 398–409.

Geman, S. and D. Geman (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica 7*, 339–373.

Green, P. J. (1995). Reversible-Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika 82*, 711–732.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika 57*, 97–109.

Heard, N. A., C. C. Holmes, and D. A. Stephens (2006). A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *Journal of the American Statistical Association 101*, 18–29.

Hoff, P. D. (2006). Model-Based Subspace Clustering. *Bayesian Analysis 1*, 321–344.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical S 20*, 50–67.

Ji, C. and S. C. Schmidler (2009). Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection. Technical report, Duke University.

Kim, S., M. G. Tadesse, and M. Vannucci (2006). Variable Selection in Clustering via Dirichlet Process Mixture Models. *Biometrika 93*, 877–893.

Lamnisos, D., J. E. Griffin, and M. F. J. Steel (2011). Adaptive Monte Carlo for Bayesian Variable Selection in Regression Models. Technical report, University of Warwick, University of Kent.

Lamnisos, D., J. E. Griffin, and M. F. J. Steel (2012). Cross-Validation Prior Choice in Bayesian Probit Regression with Many Covariates. *Statistics and Computing 22*, 359–373.

Lian, H. (2010). Sparse Bayesian Hierarchical Modeling of High-Dimensional Clustering Problems. *Journal of Multivariate Analysis 101*, 1728–1737.

Liu, J. S. (1994). The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem. *Journal of the American Statistical Association 89*, 958–966.

Liu, J. S., J. L. Zhang, M. L. Palumbo, and C. E. Lawrence (2003). Bayesian Clustering With Variable and Transformation Selections. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics*, Volume 7, pp. 249–275. Oxford University Press.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models.* New York.

Medvedovic, M. and S. Sivaganesan (2002). Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles. *Bioinformatics 18*, 1194–1206.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. Teller, and E. Teller (1953). Equation of State Calculation by Fast Computing Machines. *The Journal of Chemical Physics 21(6)*, 1087–1092.

Nobile, A. and A. T. Fearnside (2007). Bayesian Finite Mixtures With an Unknown Number of Components : The Allocation Sampler. *Statistics and Computing 17*, 147–162.

O'Hara, R. B. and M. J. Sillanpää (2009). A Review of Bayesian Variable Selection Methods : What, How and Which. *Bayesian Analysis 4*, 85–118.

Raftery, A. E. and N. Dean (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association 101*, 168–178.

Ramoni, M. F., P. Sebastiani, and I. S. Kohane (2002). Cluster Analysis of Gene Expression Dynamics. *Proceedings of the National Academy of Sciences 99*, 9121–9126.

Redner, R. A. and H. F. Walker (1984). Mixture Densitie, Maximum Likelihood and the EM Algorithm. *SIAM Review 26*, 195–239.

Richardson, S. and P. J. Green (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society Series B, 59*, 731–792.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *Annals of Applied Probability 7*, 110–120.

Roberts, G. O. and J. S. Rosenthal (2001). Optimal Scaling of Various Metropolis-Hastings Algorithms. *Statistical Science 16, 4*, 351–367.

Sha, N., M. Vannucci, P. J. Brown, M. K. Trower, and G. Amphlett (2003). Gene Selection in Arthritis Classification With Large-Scale Microarray Expression Profiles. *Comparative and Functional Genomics 4*, 171–181.

Steinley, D. and M. J. Brusco (2008). Selection of Variables in Cluster Analysis: an Empirical Comparison of Eight Procedures. *Psychometrika 73*, 125–144.

Stephens, M. (2000). Bayesian Analysis of Mixture Models With an Unknown Number of Components - an Alternative to Reversible-Jump Methods. *The Annals of Statistics 28*, 40–74.

Tadesse, M. G., N. Sha, and M. Vannucci (2005). Bayesian Variable Selection in Clustering High-Dimensional Data. *Journal of the American Statistical Association 100*, 602–617.

Yau, C. and C. C. Holmes (2011). Hierarchical Bayesian Nonparametric Mixture Models for Clustering With Variable Relevance Determination. *Bayesian Analysis 6*, 329–352.