BAYESIAN VARIABLE SELECTION IN CLUSTERING VIA DIRICHLET PROCESS MIXTURE MODELS

A Dissertation

by

SINAE KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2006

Major Subject: Statistics

BAYESIAN VARIABLE SELECTION IN CLUSTERING VIA DIRICHLET PROCESS MIXTURE MODELS

A Dissertation

by

SINAE KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Marina Vannucci
Committee Members,	Jeffrey D. Hart
	David B. Dahl
	Arul Jayaraman
Head of Department,	Simon J. Sheather

May 2006

Major Subject: Statistics

ABSTRACT

Bayesian Variable Selection in Clustering
via Dirichlet Process Mixture Models. (May 2006)
Sinae Kim, B.S., Pusan National University;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Marina Vannucci

The increased collection of high-dimensional data in various fields has raised a strong interest in clustering algorithms and variable selection procedures. In this dissertation, I propose a model-based method that addresses the two problems simultaneously. I use Dirichlet process mixture models to define the cluster structure and to introduce in the model a latent binary vector to identify discriminating variables. I update the variable selection index using a Metropolis algorithm and obtain inference on the cluster structure via a split-merge Markov chain Monte Carlo technique. I evaluate the method on simulated data and illustrate an application with a DNA microarray study. I also show that the methodology can be adapted to the problem of clustering functional high-dimensional data. There I employ wavelet thresholding methods in order to reduce the dimension of the data and to remove noise from the observed curves. I then apply variable selection and sample clustering methods in the wavelet domain. Thus my methodology is wavelet-based and aims at clustering the curves while identifying wavelet coefficients describing discriminating local features. I exemplify the method on high-dimensional and high-frequency tidal volume traces measured under an induced panic attack model in normal humans.

To my beloved Dad and Mom

&

my hubby, Taewoo

ACKNOWLEDGEMENTS

I would like to express my appreciation to Professor Marina Vannucci for being strongly supportive throughout my Ph. D. program. She encouraged me by showing her energy, positive attitude, and academic ideas which never run out. She is my academic mother even though she is way younger than to be a mother. She makes me think I want to be an advisor like her in the future. Another person I thank, is Mahlet G. Tadesse, an assistant professor from University of Pennsylvania. I could not have done this all work without her sharp intuition and help. I also appreciate her cheering me up as a friend. I would like to give a special thanks to assistant professor David B. Dahl who helped me with critical reviews and many helpful suggestions. I would also like to extend my thanks to professor Jeffrey D. Hart and assistant professor Arul Jayaraman for their suggestions and help. I might not continue to study in the department if I didn't take classes from Professor Michael T. Longnecker. His great lectures realized me why I love studying statistics. I thank him for his consistent and warmhearted consideration for students, especially for international students. I also thank Professor Choongrak Kim in Pusan National University back home. Thanks to his great lectures and encouragements in my undergraduates, I am on the track of doing research in statistics. Catherine McIntyre, my English teacher in College Station, encouraged me to find a right way in studying. She helped me to improve my poor English skill a lot as well as to have a positive attitude for life. I also thank all staffs in our department, especially Ms. Marilyn Randall who has worked on all the troubles I made with smile. I thank Henrik Schemidche for maintaining computer system stable and accessible since all my research works heavily rely on computer.

I also thank all the members in my bioinformatics group. They are real inspiration to me. I learned a lot from them. Thank you, Samiran, Michael, Lan, Quincy, Sang Han, and Jae-sik.

Kyounghwa Bae, my best friend and big sister, encourages me to keep up with my work. She is always with me when I need someone whom I cry on. I would like to give a special thank to Sang-Eui Lee. Without his support, I could not think of coming U.S.A. to live and study by myself. I wish him good luck and endless happiness. I thank all my friends in Korea and in the U.S.A, especially, Chu-Yeon, Chung-Su, Sang-Hyun, Kyoungmi, Jeongja, Deukwoo, Jason and my officemates Cynthia and Mandy. I would lose my sense of humor and laugh without their supportive e-mails and prayers. I thank my parents-in-law who never rest their prayers for Taewoo and myself, and my brothers and sisters, Cheolhong, Hannae, Haneul, and Younjeong for their prayer and support.

I also thank my best friend and beloved husband, Taewoo Chung, who always takes care of many things around me since I came here so that I could concentrate on this dissertation. He always bears with my never-ending emotional up-and-down, especially in my last year here. His love has been a strong energy for my entire study. He never complains about coming down here every weekend after 3 hour driving from Dallas, just to show his love and support for 3 years. Loving and marrying him is one of the best things that ever happened in my life.

Finally, I thank my parents who always stand beside me. I could not complete this long journey without their never-ending love, support and prayer. I can not find any word to express my deep appreciation to them. I just want to tell them how much I love them. I love you, Mom and Dad, and thanks for always being there for me.

TABLE OF CONTENTS

Page

ABSTRACT . . . iii

DEDICATION . . . iv

ACKNOWLEDGEMENTS . . . v

TABLE OF CONTENTS . . . vii

LIST OF TABLES . . . ix

LIST OF FIGURES . . . x

CHAPTER

I INTRODUCTION . . . 1

- 1.1 Motivation . . . 1
- 1.2 Dirichlet process mixture models . . . 4
- 1.3 Clustering via Dirichlet process mixture models . . . 6
- 1.4 Bayesian variable selection . . . 9
- 1.5 Outline of the dissertation . . . 11

II VARIABLE SELECTION IN CLUSTERING VIA DIRICHLET PROCESS MIXTURE MODEL . . . 13

- 2.1 Introduction . . . 13
- 2.2 Model formulation . . . 15
- 2.3 Posterior inference . . . 27
- 2.4 Discussion . . . 29

III APPLICATIONS . . . 31

- 3.1 Simulation study . . . 31
- 3.2 DNA Microarray data analysis . . . 39
- 3.3 Conclusion . . . 49

IV CASE STUDY: WAVELET-BASED BAYESIAN CLUSTER ANALYSIS OF TIDAL VOLUME CURVES . . . 50

- 4.1 Introduction . . . 50
- 4.2 Wavelet-based Bayesian clustering of functional Data . . . 52

CHAPTER Page

4.3 Case study: Panic attack 54
4.4 Discussion 66

V CONCLUSION 68

REFERENCES 72

APPENDIX A 79

APPENDIX B 83

VITA 89

LIST OF TABLES

TABLE Page

1 Some selected genes with known association to leukemia 46

LIST OF FIGURES

FIGURE Page

1 Simulated data – $\alpha = 1,\ \omega = 10/p$: Trace plot for the number of clusters. 32

2 Simulated data – $\alpha = 1,\ \omega = 10/p$: Trace plot for the number of discriminating variables. 32

3 Simulated data – $\alpha = 1,\ \omega = 10/p$: Pairwise posterior probabilities $p(c_i = c_j|\boldsymbol{X})$ for Estimation based on equation (2.18) 33

4 Simulated data – $\alpha = 1,\ \omega = 10/p$: Pairwise posterior probabilities $p(c_i = c_j|\boldsymbol{X})$ for Difference between estimates based on equation (2.18) and empirical frequencies. 34

5 Simulated data – $\alpha = 1,\ \omega = 10/p$: Estimated marginal posterior probabilities $p(\gamma_j = 1|\boldsymbol{X})$ for (a) Estimation based on equation (2.20); (b) Frequency-based estimation by (2.21); (c) Difference in probability estimates from the two approaches. 36

6 Simulated data: Prior predictive distribution of number of clusters . . 37

7 Simulated data: Sensitivity analysis for (a) $\alpha = 15$ and $\omega = 10/p$: number of clusters; (b) $\alpha = 15$ and $\omega = 10/p$: number of discriminating variables; (c) $\alpha = 1$ and $\omega = 30/p$: number of clusters; (d) $\alpha = 1$ and $\omega = 30/p$: number of discriminating variables 38

8 Heatmap of all 3,571 genes 40

9 Microarray data: Trace plots for (a) Chain 1: number of clusters; (b) Chain 1: number of discriminating variables; (c) Chain 2: number of clusters; (d) Chain 2: number of discriminating variables. 41

10 Microarray data: Pairwise posterior probabilities $p(c_i = c_j|\boldsymbol{X})$ 42

11 Microarray data: Concordance of results across MCMC chains 43

12 Microarray data: Marginal posterior probabilities $p(\gamma_j = 1|\boldsymbol{X})$ 44

FIGURE Page

13 Heatmap of 116 genes selected by method 45

14 Microarray data: Trace plots with $\alpha = 1$ for (a) Chain 2: number of clusters; (b) Chain 2: number of discriminating variables. 48

15 Raw data: 51,200 data points 59

16 Baseline-adjusted Vt: 2,048 data points 59

17 Reconstructed Vt with survived coefficients 61

18 Trace plot of number of clusters 62

19 Trace plot of number of selected wavelet coefficients 62

20 Cluster analysis during the 2nd infusion 63

21 Marginal posterior probability of $\gamma_j = 1$ 65

22 Group mean curves based on cluster allocations 66

CHAPTER I

INTRODUCTION

1.1 Motivation

The topic of this dissertation is cluster analysis of high-dimensional data. Particular interest is in situations where the cost of extracting and processing data forces researchers to generate data sets with few observations and thousands of variables. It is essential to develop statistical approaches to analyze this type of data sets, since classical statistical methods do not work well with high dimensional data. Recent examples of such data come from the field of bioinformatics. Broadly, bioinformatics refers to the science of informatics as applied to biological research. Informatics is the management and analysis of data using advanced computing techniques. Bioinformatics is particularly important as an adjunct to genomics research, because of the large amount of complex data this research generates. Yet current classical statistical methods are not adequate to meet the challenge of analyzing high-dimensional data.

In this dissertation I will propose an application of the developed methodologies to cancer genomics and, in particular, to the problem of predicting and discovering cancer subtypes. For many tumor types, it is known that there may exist unknown subtypes of the cancers. For example, some leukemia cancers have similar morphological appearances but show different responses to therapy. Thus, cancer researchers question how many subtypes of cancers exist and how to discriminate patients into these subgroups. Different subtypes of cancers may respond differently to the target

---

The format and style follow that of *Journal of the American Statistical Association.*

chemotherapy. The discovery of cancer subtypes is therefore imperative to improve treatments. This leads to the need of cluster analysis. With high-dimensional data, however, it is often not appropriate to perform clustering of the observations based on all measured variables. Clusters are often confined to a small set of variables that contain all the information to classify observations into subgroups. Noisy variables, which do not have any information to discriminate observations, may mask the true cluster structures. Thus, it is crucial to identify these informative variables in order to extract correct estimation, which in turn lead to correct cluster structures.

My ultimate goal in this dissertation is to combine these two methods, clustering and variable selection, into one model using a Bayesian approach. For clustering I will adopt a model-based clustering method via Dirichlet process mixture model, and will accomplish variable selection by introducing a latent vector to identify variables that reveal the cluster structures. The Bayesian methodologies I will develop for cluster analysis of high-dimensional data have potential for useful application in cancer genomics, particularly in the problem of finding better treatments for patients of cancer subtypes.

Dirichlet process mixture (DPM) models have been used in non-parametric Bayesian statistics to estimate density functions. The properties of the Dirichlet process enable the model to uncover clusters and determine the number of clusters. The model allows *a priori* infinite number of clusters. It also averts computationally intensive Markov chain Monte Carlo (MCMC) algorithm such as reversible jump MCMC. One may argue that the use of DPM models is still computationally intensive (for example, they may require split-merge algorithms) but it is perhaps more readily implemented. Variable selection in cluster analysis is essential with high-dimensional data. I will use a stochastic search variable selection (SSVS) method in order to identify discriminating variables. The method has similarities with the approach first introduced by George and McCulloch (1993) as a Bayesian procedure to identify informative subsets of predictors in a regression model. Bayesian variable selection methods have been widely used in regression and classification models and many contributions have appeared in the literature in recent years. Yet, variable selection in cluster analysis has only recently been studied. Unlike classification and regression models, where response variables guide the selection, in cluster analysis only the predictors are observed.

My method can be applied to a variety of data or models. In this dissertation, I will first evaluate the performance of the methodology on simulated data. This will show quite promising results. I will then apply the method to DNA microarray gene expression data. Finally, I will explore possible extensions of the proposed methodologies to the analysis of high-dimensional functional data. Functional data are discretely measured over a certain period of time. Such data are usually very noisy. I will first use wavelet shrinkage methods to remove the noise and to reduce the dimension of the data and then apply my clustering method to the survived wavelet coefficients. There are many applied fields where functional data arise. In bioinformatics, for example, this type of data is widely produced in the form of protein mass spectra or as DNA microarray gene expression measured over time. Here I choose to illustrate my method on functional data arising from the psychiatric field, and look at tidal volume data measured during panic attack experiments.

The rest of this chapter has the following sections: Section 1.2 contains a brief review of Dirichlet process and Dirichlet process mixture models. In section 1.3, I explain how cluster structures are formed via Dirichlet process mixture models. Bayesian variable selection is introduced in section 1.4. I conclude the chapter by explaining the outline of the dissertation.

1.2 Dirichlet process mixture models

Bayesian nonparametric methodologies have been developed considerably within the last decades with advances in Markov chain Monte Carlo simulation methods. Antoniak (1974) and Ferguson (1973) formalized and explored the notion of a Dirichlet process (DP). Among many authors who have endeavored to cultivate nonparametric Bayesian approaches with DP priors, many papers by Escobar (1994), MacEachern (1994), MacEachern and Müeller (1998), Neal (2000) and Escobar and West (1995) are good references for reviewing the applications of nonparametric Bayesian methods via a DP.

I will briefly explain how a DP is used in nonparametric Bayesian approaches. Let $\boldsymbol{x}_i$ be a random sample from a distribution F with parameter $\boldsymbol{\theta}_i$. In the basic Bayesian formulation, the model for the parameter $\boldsymbol{\theta}_i$ can be defined as

$$\begin{aligned} \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i &\sim G, \end{aligned} \tag{1.1}$$

and it is completed by imposing a parametric distribution form on the prior distribution G. Sometimes, however, it is not realistic to assume that the prior distribution of $\boldsymbol{\theta}_i$ is of a known form, such as multivariate normal distribution or inverse Wishart distribution. This has motivated the development of nonparametric Bayesian approaches in a hierarchical set up. One of the approaches is to introduce a Dirichlet process prior on G (Antoniak 1974; Ferguson 1983). Instead of defining a parametric form for a prior distribution of $\boldsymbol{\theta}_i$, one may avoid the restriction by assuming a random distribution G on $\boldsymbol{\theta}_i$, which is drawn from a Dirichlet process with a base distribution (or location) G_0 and a concentration parameter (or precision) α. The definition of the Dirichlet process by Ferguson (1973) can be represented as:

Definition A random probability distribution G is generated by a Dirichlet process, denoted by $G \sim \mathrm{DP}(\alpha, G_0)$ if for any partition $(A_1, \cdots, A_k)$ of the sample space, the vector of random probabilities $G(A_k)$ follows a Dirichlet distribution, i.e. $(G(A_1), \cdots, G(A_k)) \sim \mathrm{Dirichlet}(\alpha G_0(A_1), \cdots, \alpha G_0(A_k))$.

The base distribution G_0 is such that $E(G) = G_0$ and has a parametric form. The concentration parameter α measures the strength of belief in G_0. Thus if α takes a large value, a sampled G is very likely to be G_0. From the definition above, a DP is considered as a distribution function over all possible distributions. Moreover, the underlying random probability distribution G is discrete with probability one, so that the support of G consists of a countably infinite set of atoms, drawn independently from G_0. This Bayesian hierarchical model with a DP prior an be written as follows:

$$\begin{aligned} \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{1.2}$$

and it is called a Dirichlet process mixture (DPM) model. Again, a DP provides a means of placing a distribution on the space of all possible distribution functions. Thus, the support of the distribution is so large that the distribution of $\boldsymbol{\theta}_i$ is no longer restricted to lie in the set of distributions, G, as in (1.1), which can be very small portion of all distribution functions. For example, in a Bayesian parametric model, G can be assumed to have a multivariate normal distribution. But defining a DP on G allows the data to a G that is skewed, is multinomial, or departs from the parametric form G_0. Data generated from (1.2) can be partitioned according to the distinct values of the parameters because of discreteness of DP. In this formulation the DPM has a obvious interpretation as a flexible mixture model in which the number of components (clusters) is random.

There is another useful expression of a DP given by Sethuraman (1994). He provides it more explicitly in terms of a *stick-breaking construction.* Considering two infinite collections of independent random variables, $U_j \sim \text{Beta}(1, \alpha)$ and $\boldsymbol{\theta}_k \sim G_0$, for $k = 1, 2, \cdots$. If $G \sim \text{DP}(\alpha, G_0)$, then the stick-breaking representation of G is as follows:

$$\begin{aligned} \pi_k(\boldsymbol{u}) &= u_k \prod_{j=1}^{k-1} (1 - u_j) \\ G &= \sum_{k=1}^{\infty} \pi_k(\boldsymbol{u}) \ \delta(\boldsymbol{\theta}_k), \end{aligned} \tag{1.3}$$

where $\delta(\boldsymbol{\theta}_k)$ is a point mass at $\boldsymbol{\theta}_k$. The mixing proportions $\pi_k(\boldsymbol{u})$ are given by successively breaking a unit length "stick" into an infinite number of pieces. The size of each successive piece, proportional to the rest of the stick, is given by an independent draw from a $\text{Beta}(1, \alpha)$ distribution. The expression above makes clear that G is discrete with probability one. In other words, the support of G consists of a countably infinite set of atoms, drawn independently from G_0. This (1.3) leads to

$$\boldsymbol{x}_i \sim \sum_{k=1}^{\infty} \pi_k(\boldsymbol{u}) f(\cdot | \boldsymbol{\theta}_k). \tag{1.4}$$

From (1.4), it is obvious that DPM has an interpretation as a mixture model with an infinite number of mixture components.

In summary, adopting a DP prior in an hierarchical Bayesian specification of the type (1.2) leads to a DPM model. The interpretation of a DPM as a Bayesian hierarchical set-up on a finite mixture model makes it very suitable for cluster analysis.

1.3 Clustering via Dirichlet process mixture models

Mixture models have been widely used in cluster analysis to estimate cluster structures. Let me assume there are K populations mixing together with their own characteristic $\boldsymbol{\phi}_k$, which can be a scalar or a vector. Under the finite mixture models to be fitted here, each sample $\boldsymbol{x}_i$ can be viewed as coming from a population M which is a mixture of these K populations (say $M_1, \cdots, M_K$) with some proportions $p_1, \cdots, p_K$, respectively. Then the probability density function of an observation $\boldsymbol{x}$ in the population M can be represented as of the finite mixture form,

$$f(\boldsymbol{x}|\boldsymbol{\phi}) = \sum_{k=1}^{K} p_k f(\boldsymbol{x}|\boldsymbol{\phi}_k),$$

where $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$ for $k = 1, \cdots, K$. When using a mixture model for cluster analysis, one encounters the question of how many clusters K there are. In order to estimate K, many approaches have been developed based on likelihood. The details of cluster analysis using a mixture model are well-explained in McLachlan and Basford (1988) including estimation of K.

Among many other Bayesian approaches to handle cluster analysis, the use of DPM models is very attractive because of the fact that one does not need to define K a priori. DPM models can be adopted for estimating mixture distributions, which are widely-used for model-based cluster analysis. Since the realizations of a DP are discrete with probability one, these models can be viewed as countably infinite mixtures (Ferguson 1983). The discreteness implies that samples from G could have a number of ties. In other words, some $\boldsymbol{\theta}_i$s have same values with other $\boldsymbol{\theta}_l$, where $l = 1, \cdots, n$, $l \neq i$. Thus, the idea of clustering can obviously be seen here. The representation via the Pólya urn scheme, described by Blackwell and MacQueen (1973), shows the cluster formation and sample allocation. In (1.2), when G is integrated out over its prior distribution, the conditional distribution of $\boldsymbol{\theta}_i$ can be represented as following Pólya urn scheme:

(1) sample $\boldsymbol{\theta}_1 \sim G_0$

(2) sample $\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 \sim \frac{\alpha}{1+\alpha}G_0 + \frac{1}{1+\alpha}\delta(\boldsymbol{\theta}_1)$

$\vdots$

(i) sample $\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1} \sim \frac{\alpha}{i-1+\alpha}G_0 + \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\boldsymbol{\theta}_j)$

$\vdots$

(n) sample $\boldsymbol{\theta}_n|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{n-1} \sim \frac{\alpha}{n-1+\alpha}G_0 + \frac{1}{n-1+\alpha}\sum_{j=1}^{n-1}\delta(\boldsymbol{\theta}_j)$,

where $\delta(y)$ is a point mass at y. As can be seen above, it is obvious that a $\boldsymbol{\theta}_i$ has its own new value randomly selected from G_0 with a probability proportional to α, and is assigned to one of the existing values with a probability proportional to number of same values previously sampled. At the end, all sampled $\boldsymbol{\theta}_i$, $i = 1, \cdots, n$ form K clusters with $K \leq n$, and each cluster k has its distinct characteristic $\boldsymbol{\phi}_k$, $k = 1, \cdots K$ such that $\boldsymbol{\theta}_i = \boldsymbol{\phi}_k$ for a subset of index $\{i\}$ in cluster k. In other words, given K, the $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n\}$ are selected from the set $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K)$ according to a multinomial distribution. Thus, it is natural to introduce configuration c_i for i-th sample for cluster analysis. Each c_i for $i = 1, \cdots, n$ tells what cluster a sample $\boldsymbol{x}_i$ goes to. When looking at the number of unique values of estimation of $\boldsymbol{c}$, one sees the number of clusters K automatically. This scheme is definitely useful for constructing clustering idea. I note that G_0 determines the prior distribution of model-specific parameters for clusters.

The idea easily induces a cluster membership variable and conditional prior probability for configuration of each sample. As associating prior distribution of c_i with data information, I calculate posterior conditional probabilities for updating these configuration. Specific details about clustering via DPM especially conjugate case will be shown in Chapter II.

1.4 Bayesian variable selection

The development of Bayesian variable selection methods has been the focus of substantial research in recent years. Variable selection in a multiple regression model implies the selection of the covariates to be included in the model. Traditional selection procedures include AIC, BIC and C_p. However, when the number of covariates is large, a computational difficulty is encountered. In order to avoid the computational issue, more heuristic approaches have been developed that restrict attention to a smaller number of potential subsets, such as stepwise regression with forward and backward selection (Miller 1990). In the same spirit of considering small subsets of potential predictors, George and Mcullouch (1993) proposed a Bayesian approach to identify subsets of potentially good variables in a multiple regression model. Here, good variables are identified as those subsets of covariates with large posterior probability. If the number of variables, denoted p in this dissertation, is enormously large, it is not feasible to calculate the posterior probabilities of all 2^p possible subsets of variables. In order to solve this problem, George and McCullouch adopted Markov chain Monte Carlo methods that employed Gibbs sampling. These proceed to sample from the multinomial posterior distribution on the subset of possible subset choices. Subsets with high posterior probabilities can then be identified by their more frequent appearance in the Gibbs sample. George and McCullouch called the method *stochastic search variable selection* (SSVS). Here I describe their approach with more detail. The key idea is to introduce a latent binary vector to index the possible subsets of variables, say $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^t$ where $\gamma_j = 1$ if the j-th variable is selected and $\gamma_j = 0$ otherwise, for $j = 1, \cdots, p$. This indicator is used to induce a mixture prior on the regression coefficients. Let me assume a $n \times 1$ response $\boldsymbol{Y}$ be related to the $n \times p$

predictor matrix $\boldsymbol{X}$ through a model of the form

$$\boldsymbol{Y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \;\sim\; N(0, \sigma^2 I),$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients and $\boldsymbol{\epsilon}$ is a vector of measurement errors. Not all the covariates in $\boldsymbol{X}$ explain the changes in $\boldsymbol{Y}$ so that variable selection plays is needed here to identify good variables. The latent binary vector $\boldsymbol{\gamma}$ is utilized to induce mixture priors on the regression coefficients of the type

$$\beta_j \;\sim\; \gamma_j N(0, c\sigma^2) + (1 - \gamma_j) I_0,$$

where β_j indicates the j-th element of $\boldsymbol{\beta}$ and I_0 a point mass at 0. The quantity c is a hyperparameter needed to be specified. If the j-th variable is selected, i.e. $\gamma_j = 1$, then the j-th regression coefficient β_j has a normal distribution. Otherwise, if $\gamma_j = 0$ the coefficient of the corresponding variable does not appear in the model. Thus the regression model above can be represented in the following hierarchical setup

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 \quad &\sim \quad N(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 I) \\ \boldsymbol{\beta}_\gamma|\sigma^2 \quad &\sim \quad N(\mathbf{0}, c\sigma^2 I_{p_\gamma}) \\ \{\gamma_j\} \quad &\sim \quad \text{Bernoulli}(\omega), \quad \text{for } j = 1, \cdots, p. \end{aligned}$$

After integrating out all other parameters in conjugate form, the marginal posterior distribution $f(\boldsymbol{\gamma}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma})$ can be computed. This distribution contains the information relevant to the variable selection. Based on the data $\boldsymbol{Y}$, the posterior probability updates the priors on each of the 2^p possible values of $\boldsymbol{\gamma}$. Subsets of $\boldsymbol{\gamma}$ with high posterior probability $f(\boldsymbol{\gamma}|\boldsymbol{Y})$ identify the submodels supported most by the data and the prior information $f(\boldsymbol{\gamma})$. Thus, $f(\boldsymbol{\gamma}|\boldsymbol{Y})$ provides a ranking that can be used to select the more promising (or good) submodels. The SSVS by Gibbs or Metropolis sampling can then be used for posterior sampling. The SSVS is controlled by some tuning parameters such as c and ω in the model. Certain guidelines on how to choose those hyperparameters can be found in George and McCulloch (1993, 1997).

As can be seen above, in the regression model case there are response variables $\boldsymbol{Y}$ to direct the selection procedure on the predictors. A similar procedure is used in classification settings (Sha et al. 2004). In classification, the observed outcome is a categorical variable that takes one of K values identifying the group from which each sample arises. A multinomial probit model can be used to link the categorical outcome, Z, to the linear predictors, $\boldsymbol{X}$, by using a data augmentation approach, as in Albert and Chib (1993). The approach introduces a latent matrix $\boldsymbol{Y}$ where the row vector indicates the propensities of sample i to belong to one of the K classes. The model fitting in the classification setting, however, is a bit more intricate because the regression model is defined in terms of latent outcomes. The MCMC procedure needs to account for this and includes a step that updates the latent values $\boldsymbol{Y}$ from their full conditionals, for example a truncated matrix-variate t-distribution in the case of a nominal response.

Cluster analysis is a much harder problem, in that there is no response variable $\boldsymbol{Y}$ to guide the selection. Very few contributions for variable selection in cluster analysis exist in the current literature. I will review some of those in chapter II, where, in particular, I will employ the method proposed by Tadesse, Sha and Vannucci (2005). Their method makes use of the latent vector $\boldsymbol{\gamma}$ to identify variables that reveal information about the cluster structure of the observations.

1.5 Outline of the dissertation

This dissertation is organized as follows. In this chapter I have reviewed DPM and Bayesian variable selection methods, and have highlighted some of the concepts that will be central to the methodology I will propose. In Chapter II, I will introduce my method for variable selection in clustering via DPM model of high-dimensional data. In Chapter III, I will explore the performance of the methodology on simulated data and illustrate an application to leukemia cancer microarray data. I will extend methods to a clustering problem on high-dimensional functional curve data in Chapter IV and present a case study. Since functional data is usually noisy and high-dimensional, I will adopt a wavelet shrinkage method to reduce dimension and to remove noise. I will then perform variable selection on the wavelet coefficients and apply the clustering method described in Chapter II. In the final chapter I will conclude the dissertation with a summary of the research and a discussion of some of my on-going research works and future research plans.

CHAPTER II

VARIABLE SELECTION IN CLUSTERING VIA DIRICHLET PROCESS MIXTURE MODEL

2.1 Introduction

In recent years, high-dimensional data sets have become common in various areas of application. Thousands of variables are collected on a few samples, complicating inference with standard statistical approaches. Often, the goal of the analysis is to uncover the group structure of the observations and identify variables that best distinguish the different groups. A typical example is the analysis of DNA microarray data, where there is interest in discovering disease subtypes and isolating discriminating genes. The results could lead to a better understanding of the underlying biological processes and help develop targeted treatment strategies.

The practical utility of variable selection is well recognized and several methods have been developed for regression and classification models (see George and McCulloch 1993, Sha et al. 2004, among others). Few contributions have been made in the context of clustering. This is a more challenging problem since there is no observed response to guide the selection. In addition, the inclusion of unnecessary variables could complicate or mask the recovery of the clusters (see Tadesse, Sha and Vannucci 2005 for a discussion on these issues). Liu, Zhang, Palumbo and Lawrence (2003) address the problem by first reducing the dimension of the data using principal component analysis then fitting on the factors a mixture model with fixed number of clusters. They use MCMC sampling techniques to update the sample allocations and the number of factors deemed relevant for the clustering. In practice, however, the number of clusters is not known and there is often interest in evaluating the actual variables. In addition, the principal components, which are linear combinations of all variables, do not have a straightforward interpretation. Recently, Friedman and Meulman (2004) have proposed an algorithmic approach to cluster observations on separate subsets of variables. They formulate the problem in terms of distance-based clustering with weighted variables. They use heuristic search strategies to find an optimal weighting of the variables while jointly minimizing the clustering criterion. Their approach works in conjunction with hierarchical clustering, and hence does not provide inference on the number of clusters nor does it provide a measure of uncertainty for the sample allocations. Model-based approaches have also recently been proposed. Hoff (2006) adopts a mixture of Gaussian distributions where different clusters are identified by mean shifts. The model parameters are updated using Markov chain Monte Carlo (MCMC) sampling techniques and Bayes factors are computed to identify discriminating variables. Both Friedman and Meulman's and Hoff's methods allow separate subsets of variables to discriminate different groups of observations. Tadesse, Sha and Vannucci (2005) have put forward a variable selection method where latent variables are introduced to identify discriminating variables and the clustering is formulated in terms of a finite mixture of Gaussian distributions with an unknown number of components. They used a reversible jump MCMC technique to allow for the creation and deletion of clusters. Their modelling approach accounts for differences in both mean and covariance parameters across components. Unlike the procedures of Friedman and Meulman and Hoff, this approach assumes that the same subsets of variables discriminate across all components. However, the variable selection technique they adopt has the advantage of allowing flexible inference on both joint and marginal posterior distributions of the variables.

In this chapter, I build on the model of Tadesse, Sha and Vannucci (2005) by formulating the clustering in terms of an infinite mixture of distributions via Dirichlet process mixtures (DPM). Samples from a Dirichlet process are discrete with probability one and can therefore produce a number of ties. This model allows us to avoid the use of the computationally intensive reversible jump MCMC technique. The variable selection is accomplished by introducing a latent binary vector updated via MCMC. The method identifies discriminating variables and provides estimates for the number of components and the sample allocations. The chapter is organized as follows. In Section 2.2, I give details on the model formulation and the MCMC procedure. Section 2.3 describes the inference mechanism with samples generated from MCMC. Section 2.4 concludes the chapter with a brief discussion.

2.2 Model formulation

2.2.1 Clustering via Dirichlet process mixture models

Mixture models are now commonly used for cluster analysis (McLachlan and Basford 1988; Banfield and Raftery 1993) . In this approach, the data are viewed as coming from a mixture of distributions, each representing a different cluster. A long standing issue in all clustering procedures, including mixture models, is the problem of determining the number of clusters. This can be handled by defining finite mixtures with an unknown number of components. Various MCMC sampling techniques, such as the reversible jump algorithm (Richardson and Green 1997; Tadesse, Sha and Vannucci 2005) and continuous time Markov birth-death processes Stephens (2000a) have been proposed to fit this model and allow for creation and deletion of components. An alternative approach is to define mixture distributions with a countably infinite number of components. These models can be implemented by employing a Dirichlet process prior for the mixing proportions (Antoniak 1974; Ferguson 1983) . Over the last decade, various MCMC sampling methods for fitting DPM models have been developed, making these models useful in practical applications (Escobar 1994; MacEachern 1994; Escobar and West 1995; MacEachern and Müller 1998).

Let me briefly recall the main feature of a Dirichlet process mixture model as described in Chapter I. Let $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ be conditionally independent p-dimensional observations arising from a mixture of distributions $F(\boldsymbol{\theta}_i)$. The model parameters specific to individual i, $\boldsymbol{\theta}_i$, are assumed to be independent draws from some distribution, G, which in turn follows a Dirichlet process prior. This leads to the following hierarchical mixture model:

$$\begin{aligned} \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|G &\sim G \\ G &\sim DP(G_0, \alpha), \end{aligned} \tag{2.1}$$

where G_0 defines a baseline distribution for the Dirichlet process prior, such that $E[G] = G_0$, and α is a concentration parameter. The Pólya urn scheme representation of the Dirichlet process provides the basis for most computational strategies to fit this model (Blackwell and MacQueen 1973). Integrating over G allows the $\boldsymbol{\theta}_i$ to be written in terms of successive conditional distributions:

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i} \sim \frac{1}{n-1+\alpha}\sum_{k\neq i}\delta(\theta_k) + \frac{\alpha}{n-1+\alpha}G_0, \tag{2.2}$$

where $\delta(\theta_k)$ is a point mass distribution at θ_k.

Given number of clusters K, a finite mixture model is expressed as follows:

$$f(\boldsymbol{x}_i|\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k),$$

where $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0$, for $k = 1, \cdots, K$. In general Bayesian hierarchical model, the following distributions are set up. This leads to :

$$\begin{aligned} \boldsymbol{x}_i | c_i, \boldsymbol{\phi} &\sim F(\boldsymbol{\phi}_{c_i}) \\ c_i | \boldsymbol{p} &\sim \text{Discrete}(p_1, \ldots, p_K) \\ \boldsymbol{\phi}_c &\sim G_0 \\ \boldsymbol{p} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K), \end{aligned} \tag{2.3}$$

where the latent variable c_i indicates the cluster allocation of sample i and $\boldsymbol{\phi}_{c_i}$ corresponds to the identical $\boldsymbol{\theta}_i$'s. Equivalent models to (2.1) can be obtained by taking the limit as $K \to \infty$ of finite mixture models with K components above. As shown in Neal (2000), integrating over the mixing proportions $\boldsymbol{p}$ and taking $K \to \infty$ in (2.3) leads to the following prior for c_i:

$$\begin{aligned} p(c_i = c_l \text{ for some } l \neq i | \boldsymbol{c}_{-i}) &= \frac{n_{-i,k}}{n-1+\alpha} \\ p(c_i \neq c_l \text{ for all } l \neq i | \boldsymbol{c}_{-i}) &= \frac{\alpha}{n-1+\alpha}, \end{aligned} \tag{2.4}$$

where $n_{-i,k}$ is the number of $c_l = k$ for $l \neq i$. Thus, sample i is allocated to an existing cluster with probability proportional to the cluster size and it is assigned to a new cluster with probability proportional to α. As shown in Antoniak (1974), the prior probability of observing exactly k distinct clusters is given by

$$p(K = k | \alpha, n) = {}_n a_k \; \alpha^k \frac{1}{A_n(\alpha)}, \tag{2.5}$$

where the coefficients ${}_n a_k$ are the absolute values of Stirling numbers of the first kind and $A_n(x) = {}_n a_1 x + {}_n a_2 x^2 + \cdots + {}_n a_n x^n$.

If G_0 in (2.3) is a conjugate prior for F, sampling from the posterior distribution using Gibbs sampling is straightforward. I will consider a procedure where conjugacy is fully exploited as described by Neal (1992). Integrating out the model parameters $\boldsymbol{\phi}_{c_i}$ simplifies the algorithm considerably, as the latent indicators c_i will then be the only parameters to be updated. The conditional probabilities for the c_i's are then given by:

$$\begin{aligned} p(c_i = c_l \text{ for some } l \neq i | \boldsymbol{c}_{-i}, \boldsymbol{x}_i) &= b\frac{n_{-i,k}}{n-1+\alpha} \int F(\boldsymbol{x}_i; \boldsymbol{\phi}) dH_{-i,k}(\boldsymbol{\phi}) \\ p(c_i \neq c_l \text{ for all } l \neq i | \boldsymbol{c}_{-i}, \boldsymbol{x}_i) &= b\frac{\alpha}{n-1+\alpha} \int F(\boldsymbol{x}_i; \boldsymbol{\phi}) dG_0(\boldsymbol{\phi}), \end{aligned} \tag{2.6}$$

where b is the appropriate normalizing constant, $H_{-i,k}$ is the posterior distribution of $\boldsymbol{\phi}$ based on the prior G_0 and all observations $\boldsymbol{x}_l$ for which $l \neq i$ and $c_l = k$.

The Gibbs sampler and the sequential importance sampling (MacEachern, Clyde and Liu 1999), which rely on the Pólya urn-based incremental update suffer from slow mixing. Several methods have been developed to overcome this problem. One such approach is the blocked Gibbs sampler of Ishwaran and James (2001) which updates blocks of parameters. Green and Richardson (2001) have proposed using split/merge moves in the spirit of their reversible jump MCMC procedure for finite mixture models (Richardson and Green 1997). Jain and Neal (2004) and Dahl (2004) have also proposed sampling schemes that involve splitting and merging of clusters to circumvent the lack of mixing of the standard Gibbs sampler. Here, I make use of Jain and Neal's (2004) split-merge MCMC procedure. The method, which is described in Section 2.3, escapes local modes by separating or combining a group of observations based on the Metropolis-Hastings algorithm.

2.2.2 Variable selection in clustering

Variable selection in the context of clustering is inherently challenging. Unlike linear models and classification problems, where the response variable is observed and guides the selection, here the sample allocations are unknown parameters that need to be estimated. Stochastic search variable selection techniques (George and McCulloch 1993; Brown, Vannucci and Fearn 1998; among others) have successfully been used in various applications to identify informative predictors. As described in Chapter I of this dissertation, these methods introduce a latent binary vector $\boldsymbol{\gamma}$ to index all possible models and use the γ_j's to induce a mixture prior on the corresponding regression coefficients. Clustering, however, is different from a regression setting and the following adjustment is needed to define the latent indicators (Tadesse, Sha and Vannucci 2005):

$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ defines a mixture distribution} \\ \gamma_j = 0 & \text{otherwise} \end{cases} . \tag{2.7}$$

The latent vector $\boldsymbol{\gamma}$ is therefore used to directly identify variables that discriminate the different groups. I denote by $\boldsymbol{X}_{(\gamma)}$ the set of variables that define mixture distributions and by $\boldsymbol{X}_{(\gamma^c)}$ the remaining variables which favor one multivariate density across all observations.

The goal is to combine the clustering and variable selection tasks and provide a unified approach. I assume that $F(\boldsymbol{\phi}_{c_i})$ in (2.3) is an infinite mixture of Gaussian distributions with component parameters $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Thus, conditional on the discriminating variables, I have

$$\boldsymbol{x}_{i(\gamma)} | c_i = k, \boldsymbol{\phi}_k, \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_{k(\gamma)}, \boldsymbol{\Sigma}_{k(\gamma)}) \tag{2.8}$$

and with $\boldsymbol{\psi} = (\boldsymbol{\eta}, \boldsymbol{\Omega})$, the non-discriminating variables follow

$$\boldsymbol{x}_{i(\gamma^c)} | \boldsymbol{\psi}, \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\eta}_{(\gamma^c)}, \boldsymbol{\Omega}_{(\gamma^c)}). \tag{2.9}$$

The likelihood function therefore consists of the contribution from the clustering and non-clustering covariates which, assuming no correlation between the two sets of variables, is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{c},\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\psi}|\boldsymbol{X}) = \\ &(2\pi)^{-\frac{n(p-p_\gamma)}{2}}|\boldsymbol{\Omega}_{(\gamma^c)}|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_{i(\gamma^c)}-\boldsymbol{\eta}_{(\gamma^c)})^T\boldsymbol{\Omega}_{(\gamma^c)}^{-1}(\boldsymbol{x}_{i(\gamma^c)}-\boldsymbol{\eta}_{(\gamma^c)})\right\} \\ \times &\prod_{k=1}^{K}(2\pi)^{-\frac{n_k p_\gamma}{2}}|\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{n_k}{2}}\exp\left\{-\frac{1}{2}\sum_{i\in C_k}(\boldsymbol{x}_{i(\gamma)}-\boldsymbol{\mu}_{k(\gamma)})^T\boldsymbol{\Sigma}_{k(\gamma)}^{-1}(\boldsymbol{x}_{i(\gamma)}-\boldsymbol{\mu}_{k(\gamma)})\right\},\end{aligned}$$

where $p_\gamma = \sum_{j=1}^{p}\gamma_j$ and $C_k = \{i : c_i = k, i = 1, \ldots, n\}$ with cardinality n_k.

For the prior specification on $\boldsymbol{\gamma}$, I consider its elements, γ_j, to be independent Bernoulli random variables with common probability,

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p}\omega^{\gamma_j}(1-\omega)^{1-\gamma_j}, \tag{2.10}$$

where ω can be elicited as the proportion of variables expected *a priori* in the discriminating set. If further knowledge on some of the variables or their interactions is available, this information can be incorporated in the prior.

As I mentioned above I specify conjugate priors and integrate out the mean and covariance parameters. I assume, for computational convenience, independence among the non-discriminating variables and set $\boldsymbol{\Omega} = \sigma^2\boldsymbol{I}_{p\times p}$. I specify the prior distributions as follows:

$$\left.\begin{array}{rclrcl}\boldsymbol{\mu}_{k(\gamma)}|\boldsymbol{\Sigma}_{k(\gamma)} & \sim & \mathcal{N}(\boldsymbol{\mu}_{0(\gamma)}, h_1\boldsymbol{\Sigma}_{k(\gamma)}) & \boldsymbol{\eta}_{(\gamma^c)}|\boldsymbol{\Omega}_{(\gamma^c)} & \sim & \mathcal{N}(\boldsymbol{\mu}_{0(\gamma^c)}, h_0\boldsymbol{\Omega}_{(\gamma^c)}) \\ \boldsymbol{\Sigma}_{k(\gamma)} & \sim & \mathcal{IW}(\delta; \boldsymbol{Q}_{1(\gamma)}) & \sigma^2 & \sim & \mathcal{IG}(a,b)\end{array}\right., \tag{2.11}$$

where $\mathcal{IW}(\delta;\boldsymbol{Q}_1)$ is an inverse Wishart distribution with dimension p, shape parameter $\delta = n-p+1$, n degrees of freedom, and mean $\boldsymbol{Q}_1/(\delta-2)$ (Brown, 1993). $\mathcal{IG}(a,b)$ is an inverse-gamma density with mean $\frac{b}{a-1}$ and variance $\frac{b^2}{(a-1)^2(a-2)}$. Small values of δ lead to a weak prior information. I set $\delta = a = 3$, the smallest integer such that the mean and variance of the corresponding densities are defined and take $\boldsymbol{Q}_1 = \kappa_1\boldsymbol{I}_{p\times p}$.

Some care is needed in the choice of κ_1 and b. These hyperparameters need to be specified in the range of variability of the data. I found values close to the mean variance of the columns of $\boldsymbol{X}$ to yield reasonable results. For the mean parameters, I take the priors to be fairly flat over the region where the data are defined. Each element of $\boldsymbol{\mu}_0$ is set to the corresponding covariate interval midpoint. Values of h_0 and h_1 between 10 and 1,000 performed well. These data-based priors ensure that the prior distributions overlap with the likelihood and that I obtain well-behaved posterior densities. As mentioned in Richardson and Green (1997), in mixture models it is not possible to be fully non-informative and obtain proper posterior distributions. This point is also emphasized by Wasserman (2000) who proposed data-dependent priors in the context of finite mixtures. A comprehensive discussion on various prior specifications and their effects is provided in Kass and Wasserman (1996). The authors argue that the use of diffuse proper priors in complex statistical models can lead to posteriors with undesirable properties.

After integrating out the component parameters, the marginalized likelihood becomes

$$\begin{aligned} f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{c}) &= 2^{-\frac{n(p-p_\gamma)}{2}}\pi^{-\frac{np}{2}}\prod_{k=1}^{K}\left[\boldsymbol{H}_{k(\gamma)}\cdot|\boldsymbol{Q}_{1(\gamma)}|^{\frac{\delta+p_\gamma-1}{2}}\cdot\left|\boldsymbol{Q}_{1(\gamma)}+\boldsymbol{S}_{k(\gamma)}\right|^{-\frac{n_k+\delta+p_\gamma-1}{2}}\right] \\ &\times\ \boldsymbol{H}_{0(\gamma^c)}\cdot\left[\boldsymbol{S}_{0(\gamma^c)}\right]^{-(a+n/2)}, \end{aligned} \tag{2.12}$$

$$\begin{aligned} \text{where } \boldsymbol{H}_{k(\gamma)} &= (h_1 n_k+1)^{-\frac{p_\gamma}{2}}\prod_{j=1}^{p_\gamma}\frac{\Gamma\left(\frac{n_k+\delta+p_\gamma-j}{2}\right)}{\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)}, \\ \boldsymbol{H}_{0(\gamma^c)} &= (h_0 n+1)^{-\frac{p-p_\gamma}{2}}b^{a(p-p_\gamma)}\prod_{j=1}^{p-p_\gamma}\frac{\Gamma(a+n/2)}{\Gamma(a)}, \\ \boldsymbol{S}_{k(\gamma)} &= \sum_{i\in C_k}(\boldsymbol{x}_{i(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})(\boldsymbol{x}_{i(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})^T \\ &+\ \frac{n_k}{h_1 n_k+1}(\boldsymbol{\mu}_{0(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})(\boldsymbol{\mu}_{0(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})^T, \\ \boldsymbol{S}_{0(\gamma^c)} &= \prod_{j=1}^{p-p_\gamma}\left[b+\frac{1}{2}\left\{\sum_{i=1}^{n}(x_{ij(\gamma^c)}-\bar{x}_{j(\gamma^c)})^2+\frac{n}{h_0 n+1}(\mu_{0j(\gamma^c)}-\bar{x}_{j(\gamma^c)})^2\right\}\right], \end{aligned}$$

with $\bar{\boldsymbol{x}}_{k(\gamma)}$ the sample mean of cluster k, and $\bar{x}_{j(\gamma^c)}$ the j-th non-discriminating variable sample mean. The derivation of the marginalized likelihood can be found in Appendix A.

2.2.3 Model fitting

I update the variable selection index using repeated Metropolis steps and carry out inference on the cluster structure using the Jain and Neal (2004) split-merge algorithm. Our MCMC procedure iterates between the following steps:

(I) Update the latent variable selection indicator $\boldsymbol{\gamma}$ by repeating the following Metropolis step t times. A new candidate $\boldsymbol{\gamma}^{new}$ is generated by randomly choosing one of two transition moves:

  (i) Add/Delete: randomly pick one of the p indices in $\boldsymbol{\gamma}^{old}$ and change its value.

  (ii) Swap: draw independently and at random a 0 and a 1 in $\boldsymbol{\gamma}^{old}$ and switch their values.

The new candidate is accepted with probability

$$\min\left\{1, \frac{f(\boldsymbol{\gamma}^{new}|\boldsymbol{X}, \boldsymbol{c})}{f(\boldsymbol{\gamma}^{old}|\boldsymbol{X}, \boldsymbol{c})}\right\}, \tag{2.13}$$

where $f(\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{c}) \propto f(\boldsymbol{X}|\boldsymbol{\gamma}, \boldsymbol{c})p(\boldsymbol{\gamma})$. This stochastic update was suggested for model selection by Madigan and York (1995) and has been used extensively for variable selection in linear models by George and McCulloch among others, and in classification by Sha et al. (2004) . In the context of clustering, I am dealing with a more complex model where there is no observed outcome to guide the selection. Instead, the variable selection and the cluster structure evolve simultaneously. Therefore, to allow the selection to stabilize for a given cluster configuration, I repeat the Metropolis steps a number of times. In general, I found little improvement in the MCMC performance beyond 20 intermediate Metropolis steps.

(II) Update the latent sample allocation vector $\boldsymbol{c}$ using Jain and Neal's (2004) split-merge MCMC procedure. The method proceeds as follows. Start by selecting two distinct observations, i and l at random uniformly. Let $\mathcal{C}$ denote the set of observations, $k \in \{1, \ldots, n\}$, for which $k \neq i$, $k \neq l$, and $c_k = c_i$ or $c_k = c_l$.

  (1) If $\mathcal{C}$ is empty, a simple random split-merge algorithm is used:

    (a) If $c_i = c_l$, then

      (i) The component is split such that a new component $c_i^{split} \notin \{c_1, \ldots, c_n\}$ is created. The allocations for the other observations remain unchanged.

      (ii) The proposal is accepted with probability

$$a(\boldsymbol{c}^{split}, \boldsymbol{c}) = \min\left[1, \frac{q(\boldsymbol{c}|\boldsymbol{c}^{split})P(\boldsymbol{c}^{split})L(\boldsymbol{c}^{split}|\boldsymbol{X}, \boldsymbol{\gamma})}{q(\boldsymbol{c}^{split}|\boldsymbol{c})P(\boldsymbol{c})L(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{\gamma})}\right],$$

$$\text{where } \frac{q(\boldsymbol{c}|\boldsymbol{c}^{split})}{q(\boldsymbol{c}^{split}|\boldsymbol{c})} = 1, \ \frac{P(\boldsymbol{c}^{split})}{P(\boldsymbol{c})} = \alpha,$$

$$\begin{aligned} \frac{L(\boldsymbol{c}^{split}|\boldsymbol{X}, \boldsymbol{\gamma})}{L(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{\gamma})} &= \\ & \frac{\int F(\boldsymbol{x}_i; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma}) \cdot \int F(\boldsymbol{x}_l; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})}{\int F(\boldsymbol{x}_i; \boldsymbol{\phi}, \boldsymbol{\gamma}) F(\boldsymbol{x}_l; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})} \qquad (2.14) \\ = \ & \frac{(1+2h_1)^{p_\gamma/2}}{(1+h_1)^{p_\gamma}} \cdot \frac{|\boldsymbol{Q}_{1(\gamma)}|^{(\delta+p_\gamma-1)/2} \cdot |\boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{il(\gamma)}|^{(\delta+p_\gamma+1)/2}}{\left(|\boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{i(\gamma)}| \cdot |\boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{l(\gamma)}|\right)^{(\delta+p_\gamma)/2}} \\ \times \ & \prod_{j=1}^{p_\gamma} \frac{\left[\Gamma\left(\frac{\delta+p_\gamma+1-j}{2}\right)\right]^2}{\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\Gamma\left(\frac{\delta+p_\gamma+2-j}{2}\right)}, \end{aligned}$$

$$\begin{aligned}
\boldsymbol{S}_{i(\gamma)} &= (1+h_1)^{-1}(\boldsymbol{x}_{i(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})(\boldsymbol{x}_{i(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})^T, \\
\boldsymbol{S}_{il(\gamma)} &= (1+2h_1)^{-1}\left[(\boldsymbol{x}_{i(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})(\boldsymbol{x}_{i(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})^T\right. \\
&+ (\boldsymbol{x}_{l(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})(\boldsymbol{x}_{l(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})^T \\
&+ \left. h_1(\boldsymbol{x}_{i(\gamma)} - \boldsymbol{x}_{l(\gamma)})(\boldsymbol{x}_{i(\gamma)} - \boldsymbol{x}_{l(\gamma)})^T\right].
\end{aligned}$$

(b) If $c_i \neq c_l$, then

(i) c_i and c_l are merged into a single component, $\boldsymbol{c}^{merge}$.

(ii) The proposal is accepted with probability

$$a(\boldsymbol{c}^{merge}, \boldsymbol{c}) = \min\left[1, \frac{q(\boldsymbol{c}|\boldsymbol{c}^{merge})P(\boldsymbol{c}^{merge})L(\boldsymbol{c}^{merge}|\boldsymbol{X}, \boldsymbol{\gamma})}{q(\boldsymbol{c}^{merge}|\boldsymbol{c})P(\boldsymbol{c})L(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{\gamma})}\right],$$

$$\text{where } \frac{q(\boldsymbol{c}|\boldsymbol{c}^{merge})}{q(\boldsymbol{c}^{merge}|\boldsymbol{c})} = 1, \; \frac{P(\boldsymbol{c}^{merge})}{P(\boldsymbol{c})} = \frac{1}{\alpha},$$

$$\frac{L(\boldsymbol{c}^{merge}|\boldsymbol{X}, \boldsymbol{\gamma})}{L(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{\gamma})} = \frac{\int F(\boldsymbol{x}_i; \boldsymbol{\phi}, \boldsymbol{\gamma})F(\boldsymbol{x}_l; \boldsymbol{\phi}, \boldsymbol{\gamma})dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})}{\int F(\boldsymbol{x}_i; \boldsymbol{\phi}, \boldsymbol{\gamma})dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma}) \cdot \int F(\boldsymbol{x}_l; \boldsymbol{\phi}, \boldsymbol{\gamma})dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})}. \quad (2.15)$$

(2) If $\mathcal{C}$ is not empty, a restricted Gibbs sampling split-merge is used:

(a) Start by building a launch state as follows:

(i) If $c_i = c_l$, then split the component such that $c_i^{launch} \notin \{c_1, \ldots, c_n\}$ and $c_l^{launch} = c_l$.

(ii) If $c_i \neq c_l$, then $c_i^{launch} = c_i$ and $c_l^{launch} = c_l$.

(iii) For every $k \in \mathcal{C}$, set c_k^{launch} independently and at random with probability 0.5 to either c_i^{launch} or c_l^{launch}.

(iv) Perform t intermediate restricted Gibbs sampling scans to allocate each observation $k \in \mathcal{C}$ to either c_i^{launch} or c_l^{launch} using the following conditional distribution

$$p(c_k|\boldsymbol{c}_{-k}, \boldsymbol{x}_k, \boldsymbol{\gamma}) = \frac{n_{-k,c_k}A(c_k)}{n_{-k,c_i^l}A(c_i^l) + n_{-k,c_l^l}A(c_l^l)}, \quad (2.16)$$

where

$$
\begin{aligned}
A(c_i) = \int F(\boldsymbol{x}_k; \boldsymbol{\phi}, \boldsymbol{\gamma}) dH_{-k,c_i}(\boldsymbol{\phi}, \boldsymbol{\gamma}) = \\
\pi^{-p_\gamma/2} \left(\frac{h_1 n_{c_i} + 1}{h_1 n_{-k,c_i} + 1} \right)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\left(\frac{n_{c_i}+\delta+p_\gamma-j}{2}\right)}{\Gamma\left(\frac{n_{-k,c_i}+\delta+p_\gamma-j}{2}\right)} \\
\times \left| \boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{c_i(\gamma)} \right|^{-(n_{c_i}+\delta+p_\gamma-1)/2} \\
\times \left| \boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{-k,c_i(\gamma)} \right|^{(n_{-k,c_i}+\delta+p_\gamma-1)/2},
\end{aligned}
$$

$$
\begin{aligned}
\text{with } \boldsymbol{S}_{-k,c_i(\gamma)} &= \textstyle\sum_{j \neq k: c_j = c_i} \left(\boldsymbol{x}_{j(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)} \right) \left(\boldsymbol{x}_{j(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)} \right)^T \\
&+ \tfrac{n_{-k,c_i}}{h_1 n_{-k,c_i}+1} \left(\boldsymbol{\mu}_{0(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)} \right) \left(\boldsymbol{\mu}_{0(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)} \right)^T
\end{aligned}
$$

and $\boldsymbol{S}_{c_i(\gamma)}$ is defined as in equation (2.12).

The derivation of $A(c_i)$ is shown in Appendix A. Jain and Neal (2004) found that the improvement in mixing is minimal after five intermediate scans. The result from the last restricted Gibbs sampling scan constitutes the launch state for the split-merge procedure.

(b) If $c_i = c_l$, then

  (i) Let $c_i^{split} = c_i^{launch}$ and $c_l^{split} = c_l^{launch}$.

  (ii) For every observation $k \in \mathcal{C}$, perform one final Gibbs sampling scan from $\boldsymbol{c}^{launch}$ to set c_k^{split} to either c_i^{split} of c_l^{split} using equation (2.16).

  (iii) The allocation for observations $k \notin \mathcal{C} \cup \{i, l\}$ remains unchanged, $c_k^{split} = c_k$.

  (iv) Evaluate the proposal by the Metropolis-Hastings acceptance probability $a(\boldsymbol{c}^{split}, \boldsymbol{c})$, where $q(\boldsymbol{c}^{split}|\boldsymbol{c})$ is obtained by computing the Gibbs sampling transition probability from $\boldsymbol{c}^{launch}$ to $\boldsymbol{c}^{split}$.

(c) If $c_i \neq c_l$, then

  (i) Let $c_i^{merge} = c_l$ and $c_l^{merge} = c_l$.

  (ii) For every observation $k \in \mathcal{C}$, let $c_k^{merge} = c_l$.

  (iii) The allocation for observations $k \notin \mathcal{C} \cup \{i, l\}$ remains unchanged, $c_k^{merge} = c_k$.

  (iv) The proposal is accepted with probability $a(\boldsymbol{c}^{merge}, \boldsymbol{c})$, where $q(\boldsymbol{c}|\boldsymbol{c}^{merge})$ is the product over $k \in \mathcal{C}$ of the probabilities of setting each c_k in the original split state to its value in the launch state.

One iteration is completed after performing a full Gibbs sampling scan and updating all sample allocations $c_i,\ i = 1, \ldots, n$.

The split-merge algorithm helps improve the mixing of the MCMC sampler, which is a typical problem in fitting mixture models. The problem here is further aggravated by the inclusion of variable selection. In cases where the sampler still exhibits poor performance, getting stuck at a local mode and not accepting the proposed split-merge moves, a tempering scheme can be introduced. One such approach is the parallel tempering algorithm (Geyer, 1991). A series of distributions that interpolate between the distribution of interest and a distribution from which sampling is easier are defined, such that $f_t(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{\gamma}) = f(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{\gamma})^{1/T_t}$ for $t = 1, \ldots, T$. The procedure consists of the following steps:

1. parallel scan: for each chain with equilibrium distribution $f_t(.)$, $\boldsymbol{c}^{old}(T_t)$ is updated to $\boldsymbol{c}^{new}(T_t)$ as described above.

2. state exchange: two neighboring chains, T_t and $T_{t'}$, are randomly chosen and an attempt is made to swap $\boldsymbol{c}^{new}(T_t)$ with $\boldsymbol{c}^{new}(T_{t'})$. This update is accepted with probability

$$\min\left\{1, \left[\frac{f(\boldsymbol{c}^{new}(T_{t'})|\boldsymbol{X},\boldsymbol{\gamma})}{f(\boldsymbol{c}^{new}(T_t)|\boldsymbol{X},\boldsymbol{\gamma})}\right]^{\left(\frac{1}{T_t}-\frac{1}{T_{t'}}\right)}\right\}.$$

2.3 Posterior inference

The MCMC output can be used to draw inference on the cluster structure and on the variable selection vector $\boldsymbol{\gamma}$. Different inference strategies can be adopted. A common approach for estimating discrete marginal posterior probabilities uses empirical relative frequencies. Here, I have chosen to also approximate these distributions by summing the posterior probabilities over MCMC scans that correspond to the configuration of interest. I report the inference using both my approach and the frequency-based estimates, which give concordant results.

2.3.1 Inference on $\boldsymbol{c}$

For inference on the cluster structure, a commonly used estimate is the maximum *a posteriori* (MAP) sample allocation vector, which corresponds to the configuration with highest conditional posterior probability among those drawn by the MCMC sampler:

$$\widehat{\boldsymbol{c}} = \underset{1\leq t\leq M}{\operatorname{argmax}}\, p(\boldsymbol{c}^{(t)}|\boldsymbol{X},\widehat{\boldsymbol{\gamma}}), \tag{2.17}$$

where $\widehat{\boldsymbol{\gamma}}$ is the set of variables selected based on the marginal posterior probabilities of γ_j from equation (2.20).

Alternatively, one can estimate the latent cluster assignments by identifying the configuration with largest posterior probability, $\hat{\boldsymbol{c}} = \operatorname{argmax} p(\boldsymbol{c} = c|\boldsymbol{X})$, where

$$\begin{aligned} p(\boldsymbol{c} = c|\boldsymbol{X}) &= \int p(\boldsymbol{c} = c, \boldsymbol{\gamma}|\boldsymbol{X}) \\ &\propto \int f(\boldsymbol{X}|\boldsymbol{c} = c, \boldsymbol{\gamma}) p(\boldsymbol{c} = c) p(\boldsymbol{\gamma}) d\gamma \\ &\approx \sum_{t:\boldsymbol{c}=c} f(\boldsymbol{X}|\boldsymbol{c}^{(t)}, \boldsymbol{\gamma}^{(t)}) p(\boldsymbol{c}^{(t)}) p(\boldsymbol{\gamma}^{(t)}), \end{aligned}$$

with t indexing the MCMC iterations. The prior probability of a particular configuration $\boldsymbol{c} = c$ with D distinct mixture components, each of size n_k, $(k = 1, \cdots, D)$, is given by

$$p(\boldsymbol{c} = c) = \frac{\alpha^D \prod_{k=1}^{D} (n_k - 1)!}{\prod_{i=1}^{n} (\alpha + i - 1)}.$$

I also investigate another estimator that relies on posterior pairwise probabilities

$$\begin{aligned} p(c_i = c_j|\boldsymbol{X}) &= \int p(c_i = c_j, \boldsymbol{c}_{-(i,j)}, \boldsymbol{\gamma}|\boldsymbol{X}) d\boldsymbol{\gamma} d\boldsymbol{c}_{-(i,j)} \\ &\propto \int f(\boldsymbol{X}|c_i = c_j, \boldsymbol{c}_{-(i,j)}, \boldsymbol{\gamma}) p(c_i = c_j, \boldsymbol{c}_{-(i,j)}) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma} d\boldsymbol{c}_{-(i,j)} \\ &\approx \sum_{t:c_i=c_j} p(\boldsymbol{X}|\boldsymbol{c}^{(t)}, \boldsymbol{\gamma}^{(t)}) p(\boldsymbol{c}^{(t)}) p(\boldsymbol{\gamma}^{(t)}), \end{aligned} \tag{2.18}$$

where $\boldsymbol{c}_{-(i,j)}$ is the vector $\boldsymbol{c}$ without the i^{th} and j^{th} elements. With a sample size n there are $\binom{n}{2}$ such pairwise posterior probabilities, which can be viewed as entries of a symmetric $n \times n$ similarity matrix. An approach proposed by Dahl (2006) which he refers to as least-squares clustering, estimates the cluster structure by forming an association matrix at every MCMC iteration. Each cell of the association matrix takes value 1 if the corresponding row and column elements are allocated to the same cluster and 0 otherwise. The sum of absolute deviations between the entries of the association matrix and those of the similarity matrix is then calculated for each MCMC output, and the configuration that minimizes the quantity is considered.

2.3.2 Inference on $\boldsymbol{\gamma}$

Inference on variables that discriminate between the different groups can be done either through the joint posterior distribution of $\boldsymbol{\gamma}$ or through the marginal posterior distributions of its elements. The former selects variables based on

$$\widehat{\boldsymbol{\gamma}} = \underset{1 \leq t \leq M}{\operatorname{argmax}}\, p(\boldsymbol{\gamma}^{(t)} | \boldsymbol{X}, \widehat{\boldsymbol{c}}), \tag{2.19}$$

where $\widehat{\boldsymbol{c}}$ is the sample allocation estimated via equation (2.18). The latter identifies the variables with largest marginal posterior probabilities

$$\begin{aligned} p(\gamma_j = 1 | \boldsymbol{X}) &= \int p(\gamma_j = 1, \boldsymbol{\gamma}_{(-j)} | \boldsymbol{X}, \boldsymbol{c}) d\boldsymbol{c} d\boldsymbol{\gamma}_{(-j)} \\ &\propto \int f(\boldsymbol{X} | \boldsymbol{c}, \gamma_j = 1, \boldsymbol{\gamma}_{(-j)}) p(\boldsymbol{c}) p(\boldsymbol{\gamma}) d\boldsymbol{c} d\boldsymbol{\gamma}_{(-j)} \\ &\approx \sum_{t: \gamma_j = 1} f(\boldsymbol{X} | \boldsymbol{c}^{(t)}, \boldsymbol{\gamma}^{(t)}) p(\boldsymbol{c}^{(t)}) p(\boldsymbol{\gamma}^{(t)}), \end{aligned} \tag{2.20}$$

where $\boldsymbol{\gamma}^{(t)}$ corresponds to the vector $\boldsymbol{\gamma}$ in the model visited at the t-th iteration.

The alternative way to estimate $\boldsymbol{\gamma}$ is to use the simple frequency,

$$p(\gamma_j = 1 | \boldsymbol{X}, \hat{\boldsymbol{c}}) = \frac{1}{M} \sum_{t=1}^{M} p(\gamma_j^{(t)} = 1 | \boldsymbol{X}, \hat{\boldsymbol{c}}). \tag{2.21}$$

The two (2.20) and (2.21) give the same results of estimation. I confirm this in the chapter III.

2.4 Discussion

I have proposed a method for simultaneously uncovering cluster structure among observations and selecting discriminating variables in high-dimensional data. This approach uses model-based clustering defined via Dirichlet process mixture priors, which allow for an infinite mixture of distributions. The variable selection is accomplished by introducing a latent binary vector updated via MCMC sampling techniques. In the next chapter I will show that the methods perform well on simulated data. I will also present an example with DNA microarray data.

The use of infinite mixture models is an attractive alternative to finite mixture models, which requires a dimension jumping technique to create and delete clusters. With the DPM models, the number of components is influenced by the sample size n and the hyperparameter α. The creation and deletion of clusters is naturally taken care of in the process of updating the sample allocations.

Here, I have put forward a couple of approaches for estimating the sample allocations. One could also draw inference conditional on a fixed number of clusters, say for instance, conditioning on the value most frequently visited by the MCMC sampler. This, however, has the limitation of using only a subset of the MCMC output. In addition, with the Gibbs sampling update adopted here, a label switching problem arises since the likelihood is invariant under permutation of the component labels. This problem can be handled using Stephen's relabeling algorithm, where the MCMC output is post-processed to minimize an appropriate loss function Stephens (2000b). Alternative posterior estimators can also be obtained using the Rao-Blackwellisation method or by using decision theoretic approaches.

CHAPTER III

APPLICATIONS

I assess the performance of the methodology introduced in chapter II using a simulated data and illustrate its application on a real data set from a DNA microarray study.

3.1 Simulation study

I first investigate the performance of the methodology using simulated data. I generate a dataset of 15 observations and 1000 variables, where a set of 20 variables are chosen to separate the observations into four components

$$\begin{aligned} x_{ij} \sim\ & I_{\{1\leq i\leq 4\}}N(\mu_1,\sigma_1^2)+I_{\{5\leq i\leq 7\}}N(\mu_2,\sigma_2^2)+I_{\{8\leq i\leq 13\}}N(\mu_3,\sigma_3^2) \qquad (3.1)\\ & +I_{\{14\leq i\leq 15\}}N(\mu_4,\sigma_4^2),\ i=1,\ldots,15,\ j=1,\ldots,20, \end{aligned}$$

where $I_{\{.\}}$ is the indicator function equal to 1 if the condition is satisfied. Thus, the first 4 observations are generated from one group, the next 3 come from the second group, the next 6 are in the third group, and the last 2 fall in the fourth group. The component parameters μ_k and $\sigma_k^2,\ k=1,\ldots,4$, are randomly chosen from [-5,5] and [0.01,1] respectively. The remaining 980 variables, which do not separate the samples into clusters, are drawn from a standard normal density.

I choose the hyperparameters h_1 and κ_1 such that $h_1 \times \kappa_1$ is close to the mean of the empirical variances from the p variables. I set $h_1 = 1000$ and found the results to be quite robust for values of κ_1 in the range $[5\times 10^{-4}, 2\times 10^{-3}]$. For the non-discriminating variables, I choose b equal to the mean of the variances and found

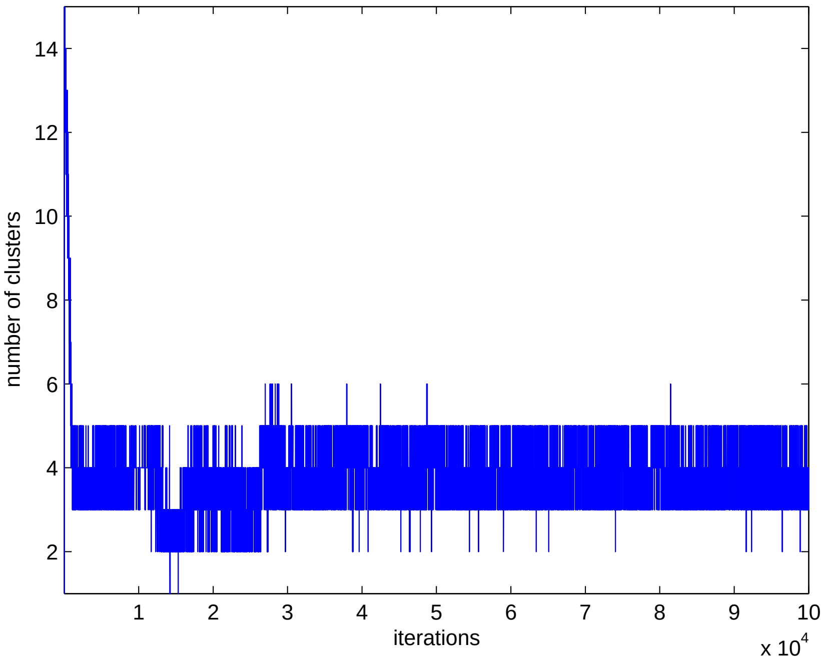


Figure 1: Simulated data – $\alpha = 1,\ \omega = 10/p$: Trace plot for the number of clusters.

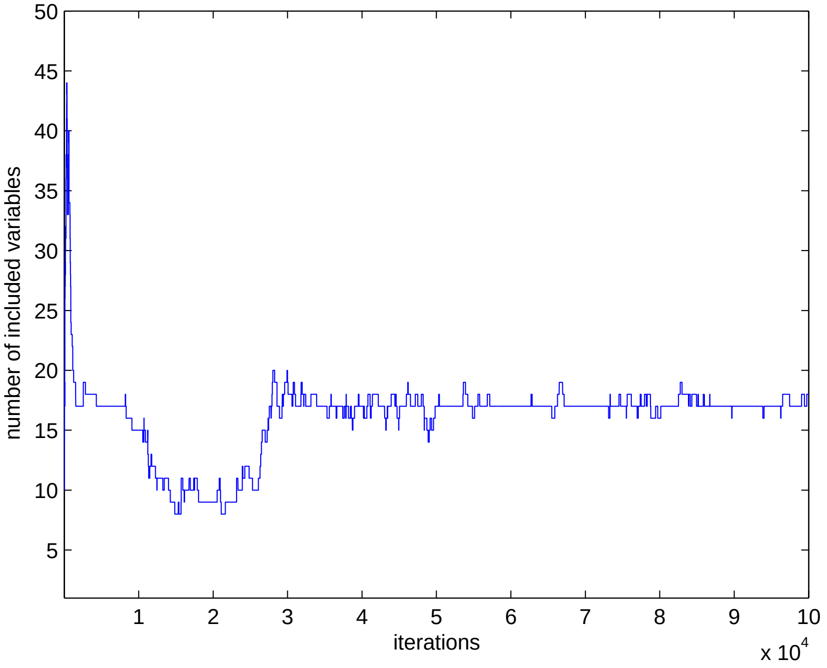


Figure 2: Simulated data – $\alpha = 1,\ \omega = 10/p$: Trace plot for the number of discriminating variables.

h_0 values between 10 and 100 to perform well. I report here the results for $\alpha = 1$, $\delta = a = 3$, $\kappa_1 = 7 \times 10^{-4}$, $h_0 = 100$, $b = 0.2$ and $\omega = 10/p$. I started an MCMC chain from a vector $\boldsymbol{\gamma}$ with 10 randomly selected elements set to 1 and each observation in a separate cluster. I ran 100,000 iterations and used the first 40,000 as burn-in. At each MCMC iteration, I performed 20 repeated Metropolis steps to update $\boldsymbol{\gamma}$ and 3 restricted Gibbs scans with one final Gibbs sampling to update $\boldsymbol{c}$. I also used the parallel tempering algorithm with two temperature ladders to further improve the mixing of the sampler. The temperatures were chosen such that the acceptance rates for exchanges between neighboring chains are between 0.5 and 0.7.

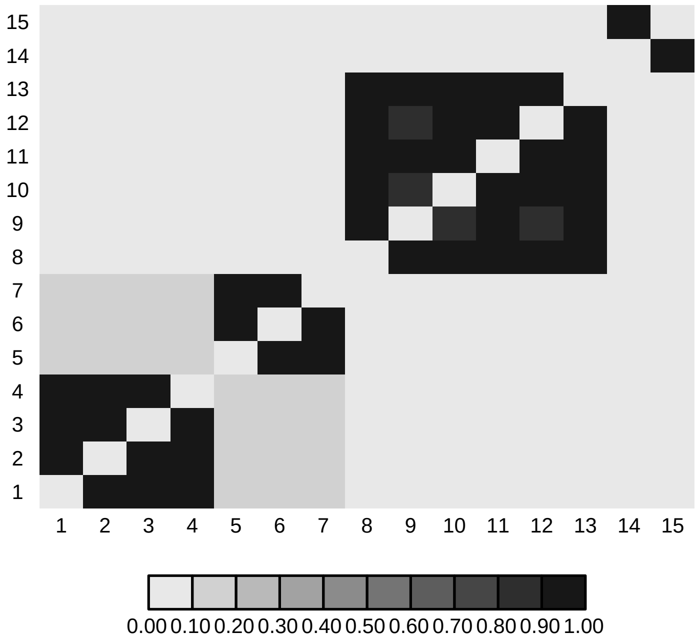


Figure 3: Simulated data – $\alpha = 1$, $\omega = 10/p$: Pairwise posterior probabilities $p(c_i = c_j|\boldsymbol{X})$ for Estimation based on equation (2.18)

Figures 1 and 2 show respectively the trace plots for the number of clusters and the number of discriminating variables. The MCMC sampler stabilized quickly around models with 3 to 5 clusters and 15 to 20 discriminating variables. I estimated the cluster allocations as described in Section 3. The posterior sample allocations estimated using equation (2.17) favored 5 components with the last two observations assigned to separate clusters. The allocations obtained using the pairwise probability estimates of equation (2.18) and the least-square clustering algorithm perfectly matched the true cluster structure.

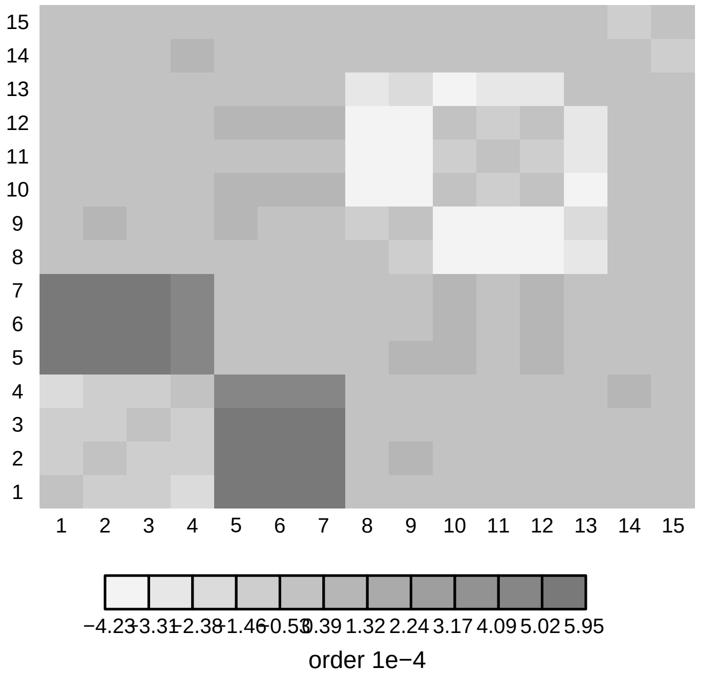


Figure 4: Simulated data – $\alpha = 1$, $\omega = 10/p$: Pairwise posterior probabilities $p(c_i = c_j|\boldsymbol{X})$ for Difference between estimates based on equation (2.18) and empirical frequencies.

Figure 3 displays the pairwise posterior probabilities, $p(c_i = c_j|\boldsymbol{X})$, of allocating observations i and j to the same cluster. The groupings used to simulate the data are successfully identified. I found identical results using the frequency-based estimates. Figure 4 plots the differences between the pairwise posterior probability estimates using the two approaches. They are in the order of 10^{-4} suggesting good concordance between the two estimators.

For the variable selection, I ordered the visited vectors $\boldsymbol{\gamma}^{(t)}$ according to their posterior probabilities and identified the "best" subset as the $\widehat{\boldsymbol{\gamma}}$ from equation (2.19). This vector contained 17 variables, all of which are among the 20 discriminating covariates used to simulate the data. I also looked at the marginal posterior probabilities, $p(\gamma_j = 1|\boldsymbol{X})$, which are displayed in Figure 5(a). The x-axis in this plot corresponds to the variable indices and the spikes indicate variables that have high posterior probabilities. The same 17 variables were selected at a marginal probability threshold of 0.7. I also report the corresponding frequency-based estimates by (2.21) in Figure 5(b). Again, I note that the same variables are selected with comparable marginal probability estimates. This concordance can also be seen in Figure 5(c), which shows for each variable the difference in marginal posterior probability estimates from the two approaches.

I investigated the sensitivity of the results to the choice of α and ω, which respectively influence the number of clusters and the number of selected variables. In general, I found the results to be quite robust to the values of these hyperparameters. Here, I report the results for two different choices of each parameter. I considered $\alpha = 1$ and $\alpha = 15$, which is equal to the sample size. As shown in equation (2.5), the number of clusters is defined *a priori* by the sample size n in the data and the choice of the hyperparameter α.

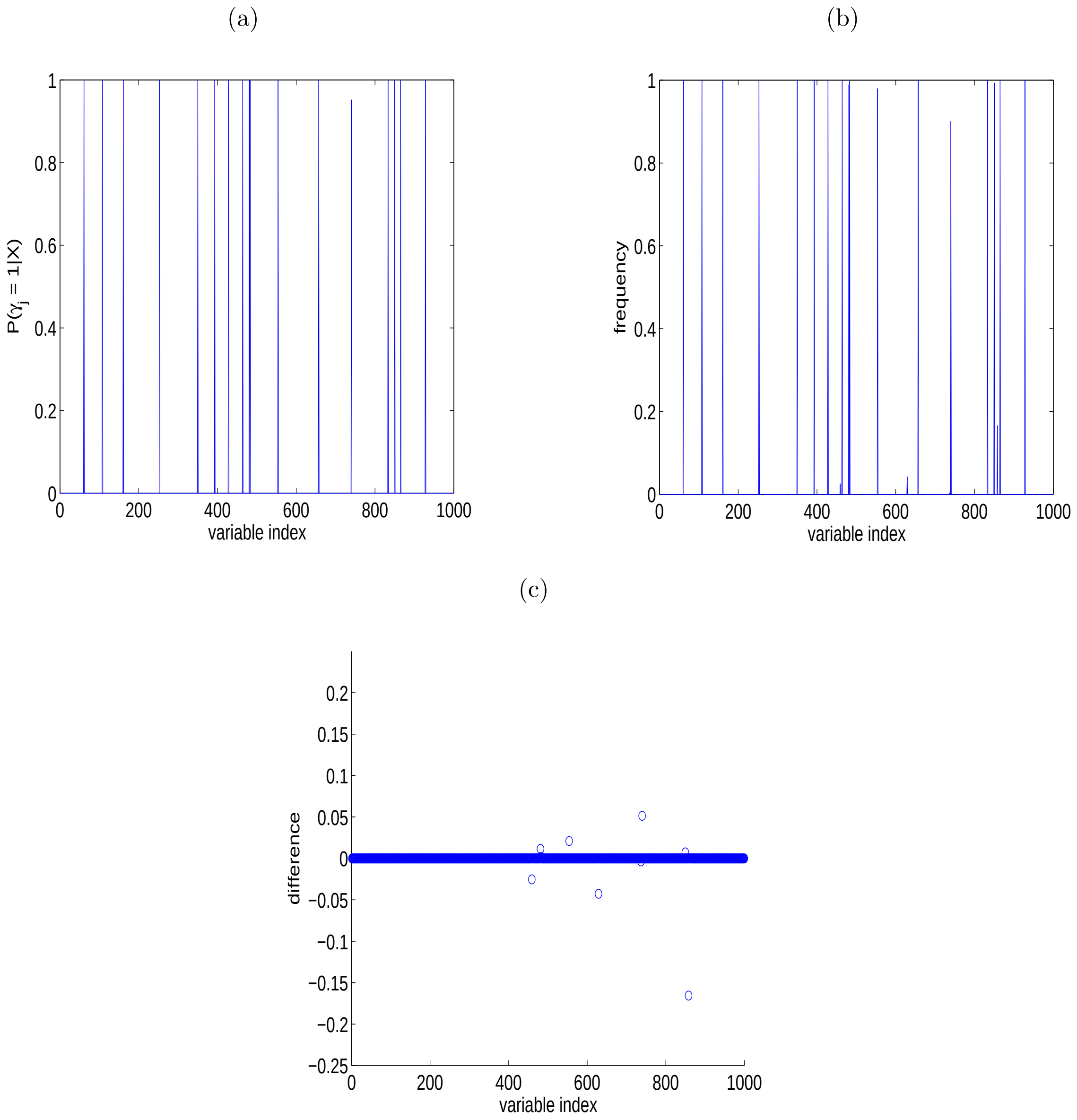

(a)
(b)
(c)

Figure 5: Simulated data – $\alpha = 1,\ \omega = 10/p$: Estimated marginal posterior probabilities $p(\gamma_j = 1|\boldsymbol{X})$ for (a) Estimation based on equation (2.20); (b) Frequency-based estimation by (2.21); (c) Difference in probability estimates from the two approaches.

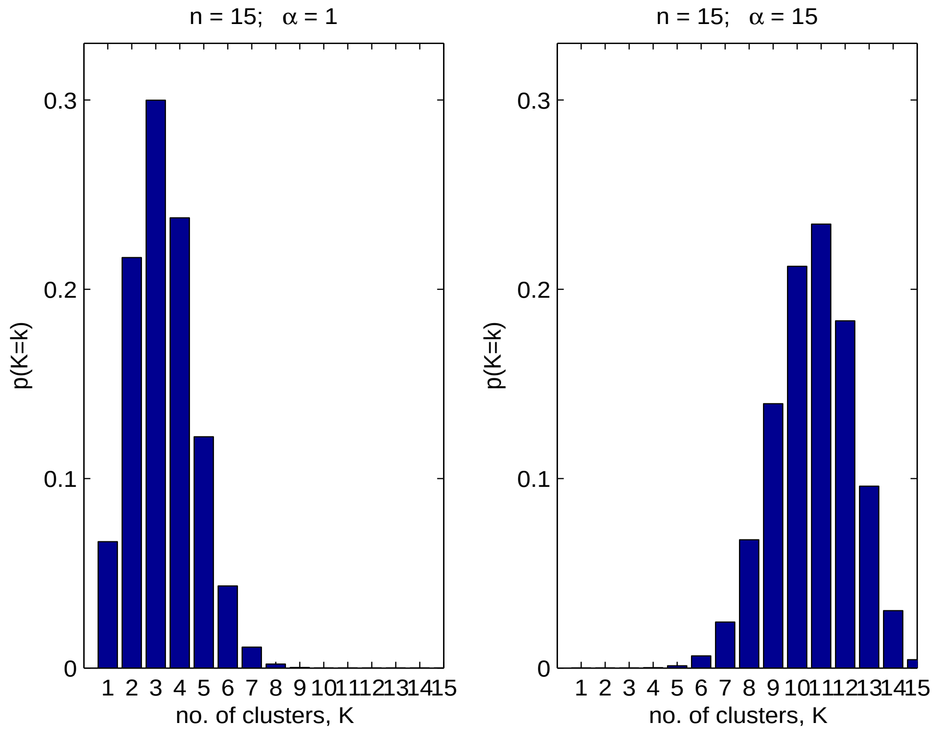


Figure 6: Simulated data: Prior predictive distribution of number of clusters

Figure 6 displays the prior predictive distribution of the number of clusters in our simulated example for the two values of α considered here. With $\alpha = 1$, the prior distribution of the number of components is concentrated between 1 and 6, whereas with $\alpha = 15$ between 7 and 14 clusters are expected *a priori*. For the variable selection hyperparameter, I chose $\omega = 10/p$ and $\omega = 30/p$. The corresponding trace plots are provided in Figure 7. For these choices, the inference on both the cluster structure and the selected variables are similar to those described above for $\alpha = 1$ and $\omega = 10/p$. The same 15 discriminating variables are selected and the four clusters are successfully identified. I note, however, that a larger value of α makes the sampler visit models with more components, although there is still strong support for models with 3 to 5 clusters (see Figure 7(a)). It appears from Figure 7(b) that a larger value of α also affects the mixing of the sampler in terms of the variable selection.

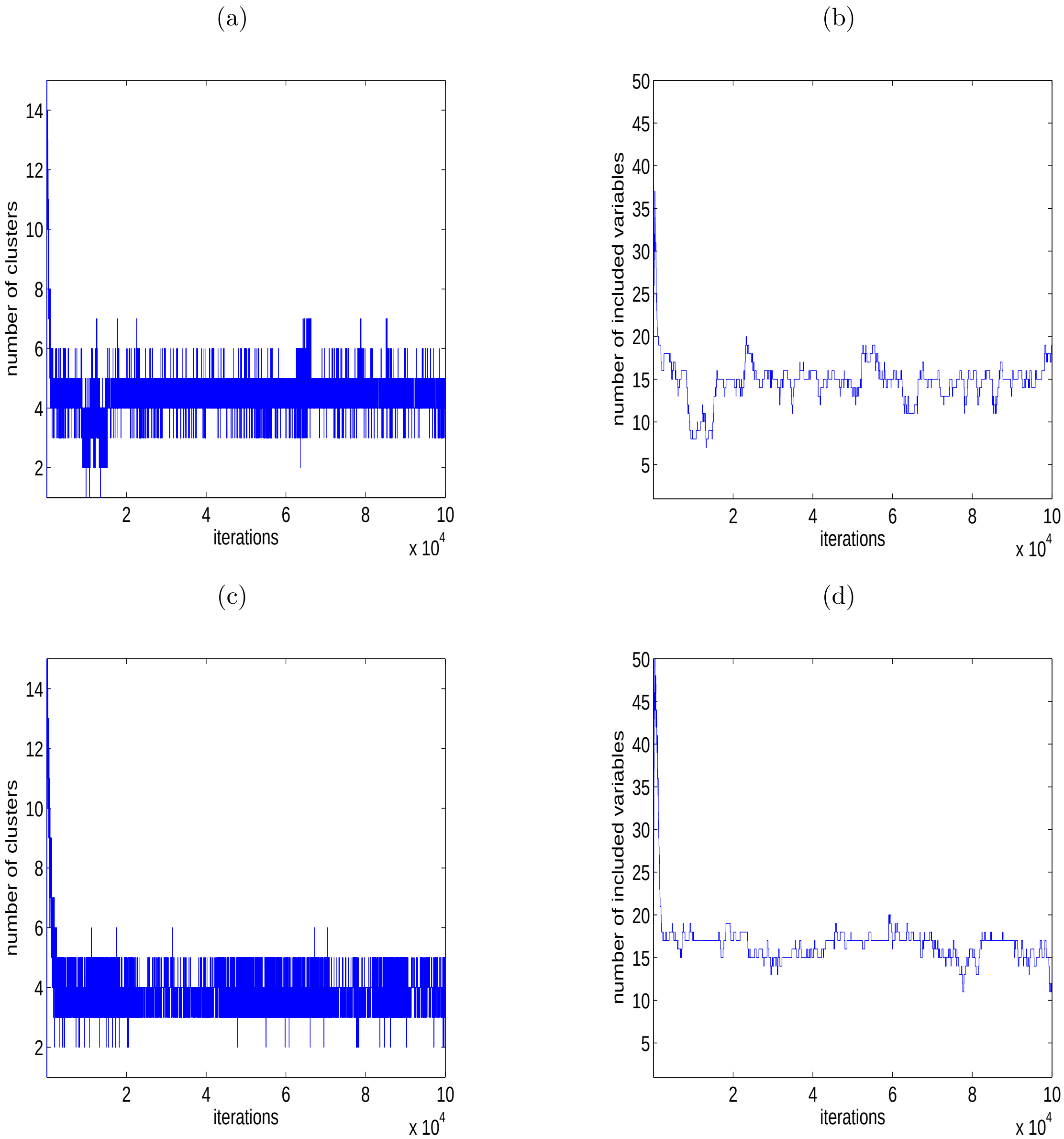


Figure 7: Simulated data: Sensitivity analysis for (a) $\alpha = 15$ and $\omega = 10/p$: number of clusters; (b) $\alpha = 15$ and $\omega = 10/p$: number of discriminating variables; (c) $\alpha = 1$ and $\omega = 30/p$: number of clusters; (d) $\alpha = 1$ and $\omega = 30/p$: number of discriminating variables

This is not surprising since the cluster structure and the variable selection evolve simultaneously.

This simulated data is identical to the one used in Tadesse, Sha and Vannucci (2005), where a finite mixture model with reversible jump MCMC technique was used to infer the cluster structures. In that paper, a detailed analysis using Friedman and Meulman's (2004) COSA algorithm, which performs variable selection in the context of hierarchical clustering, was also presented. Excellent results were obtained with the Bayesian reversible jump method, which fully recovered the true cluster structure and selected 17 of the good variables. Using the COSA approach, neither the single, average or complete linkage options for the hierarchical clustering were able to recover the true grouping of the data. The average linkage performed slightly better, being able to recover one of the clusters and identifying a set of variables that contained 16 of the discriminating covariates. This performance of COSA could be due to the fact that the method is designed to find clusters for which the discriminating variables have small variance.

3.2 DNA Microarray data analysis

A typical application where clustering has become a common task is the analysis of DNA microarray data, where thousands of gene expression levels are monitored on a few experimental units. This has revived interest in both distance-based and model-based clustering methods. For example, Medvedovic and Sivaganesan (2002) used DPM models to cluster genes with similar expression patterns. Our goal here is different. I want to uncover subclasses among the experimental units and identify genes that best discriminate between the different groups. This could help identify disease subtypes and understand some of the heterogeneity in treatment outcome for

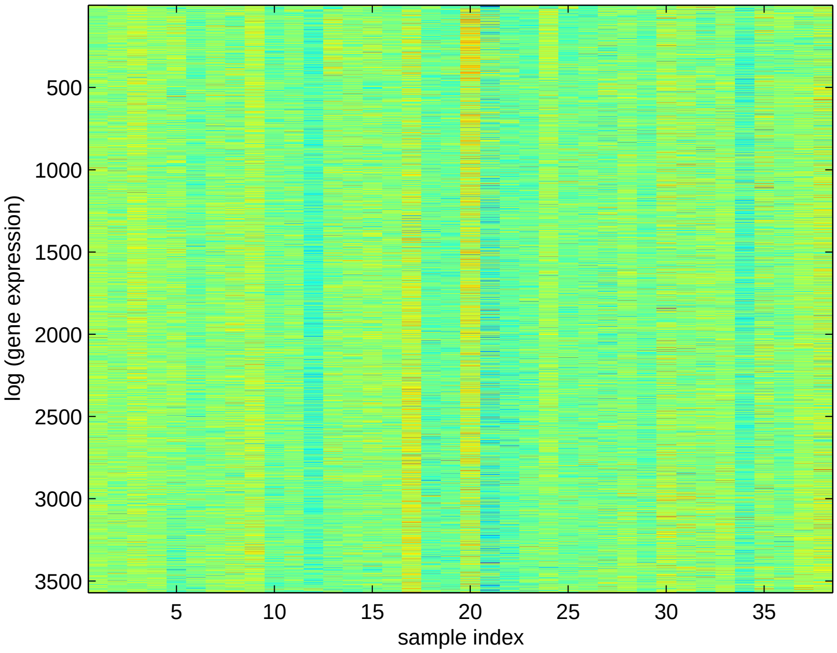


Figure 8: Heatmap of all 3,571 genes

patients receiving similar diagnoses. I illustrate our methodology using the widely analyzed leukemia data of Golub et al. (1999) and focus on the 38 patients from the training set. I followed the same pre-processing as other investigators who have analyzed the data (Dudoit, Frdlyand and Speed 2002) . Expression measures beyond the threshold of reliable detection were truncated at 100 and 16,000, and probe sets with intensities such that $\max / \min \leq 5$ and $\max - \min \leq 500$ were removed. This left 3,571 genes for analysis. The expression readings were log-transformed and each variable was rescaled by its range. Figure 8 shows a heatmap with all 3,571 genes for the analysis. As can be seen in the heatmap, it is unfeasible to discover cluster structures because of many unnecessary genes.

I chose the hyperparameters using similar guidelines to those of the simulated example. I ran MCMC chains with α set to 1 and 38, which is same as the sample size. The other hyperparameters were taken to be $\delta = a = 3$, $h_0 = 100$, $h_1 = 10$, $\kappa_1 = 0.06$,

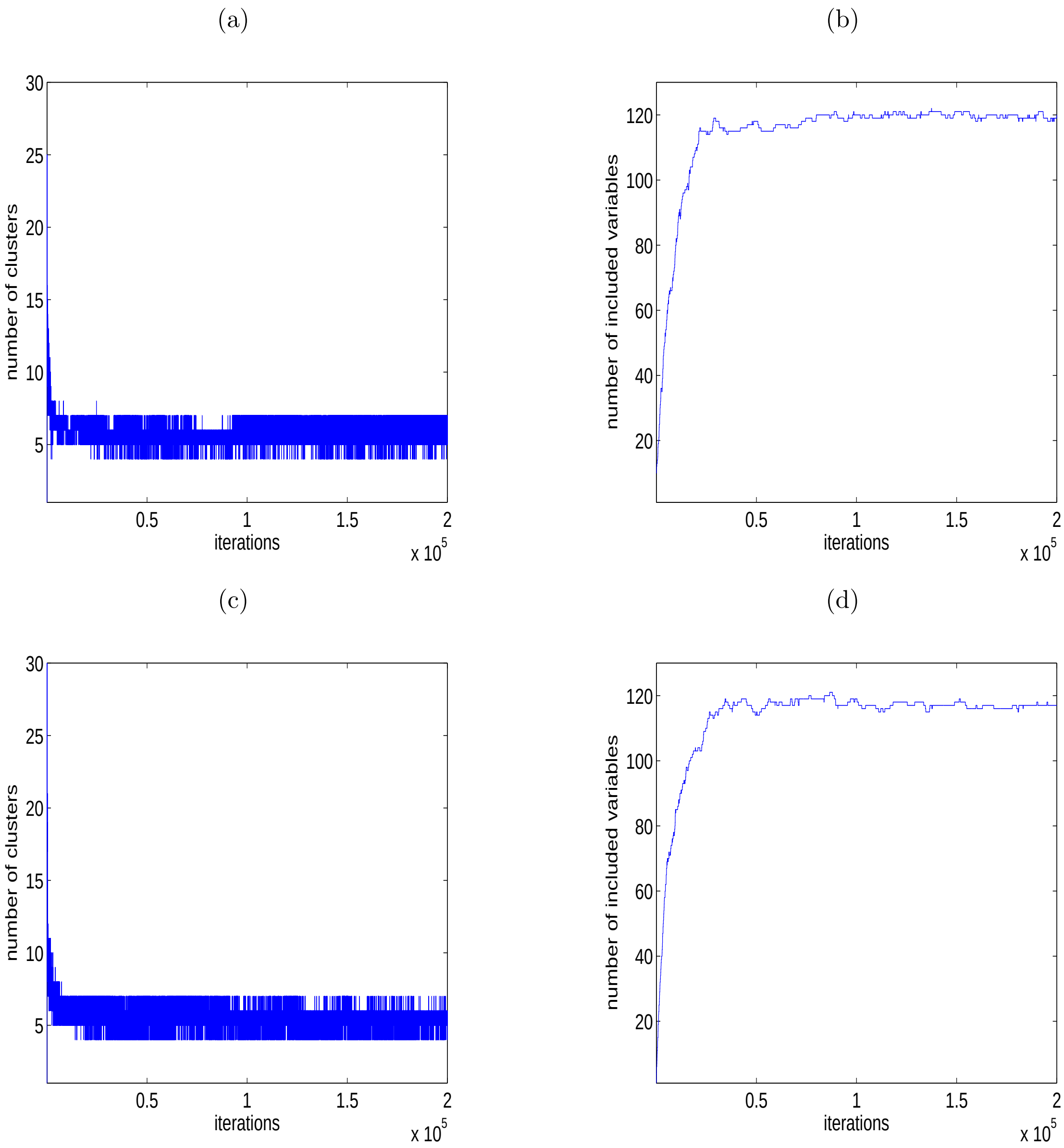


Figure 9: Microarray data: Trace plots for (a) Chain 1: number of clusters; (b) Chain 1: number of discriminating variables; (c) Chain 2: number of clusters; (d) Chain 2: number of discriminating variables.

$b = 0.1$ and $\omega = 20/p$ under the guideline suggested in Section 3.6. For both values of α, I ran two MCMC chains with different initial models: (1) all elements of $\boldsymbol{\gamma}$ set to 0 except for 10 randomly chosen γ_j's; (2) a single randomly chosen γ_j set to 1. In all cases, the sampler was started with all observations assigned to one cluster and 200,000 iterations were run with the first 100,000 used as burn-in.

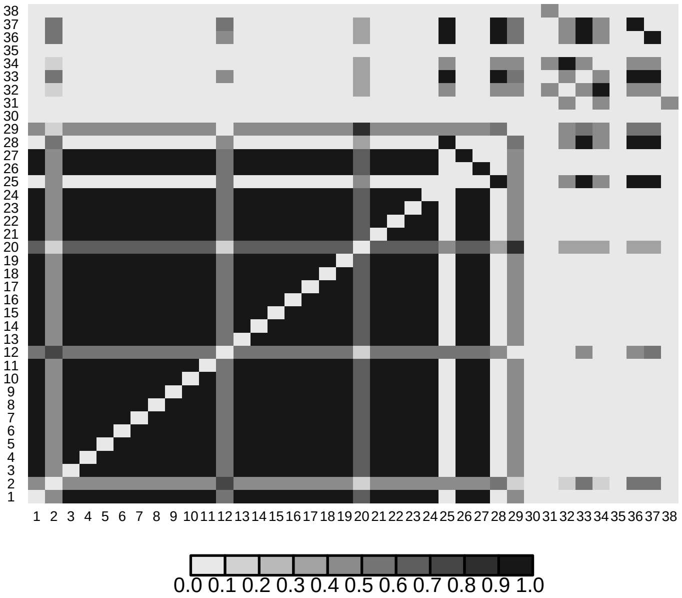


Figure 10: Microarray data: Pairwise posterior probabilities $p(c_i = c_j|\boldsymbol{X})$

Figure 9 gives the summary trace plots for the number of clusters and the number of discriminating variables using $\alpha = 38$. For both chains, the sampler mixed well mostly visiting models with 4 to 7 components. As for the number of variables, the chains stabilized near models with 120 discriminating variables. For posterior inference, I pooled the output from the two MCMC chains by taking the union of the sets of visited models. The sample allocation estimates using the different approaches presented in Section 3 were as follows:

$$\text{using (2.17) } \widehat{\boldsymbol{c}} = (\underbrace{1,2,1,\ldots,1,1,1,\ldots,1,1,1,1}_{ALL},\underbrace{2,1,4,5,3,2,3,6,2,2,7}_{AML})$$

$$\text{using (2.18) } \widehat{\boldsymbol{c}} = (\underbrace{1,2,1,\ldots,1,2,1,\ldots,1,2,1,1}_{ALL},\underbrace{2,1,4,5,3,2,3,6,2,2,7}_{AML})$$

The allocations based on the posterior pairwise probability estimates from equation (2.18) and the empirical frequencies from the MCMC output were identical. Figure 10 displays a heatmap of the pairwise posterior probabilities, $p(c_i = c_j|\boldsymbol{X})$. The first 27 indices correspond to the acute lymphoblastic leukemia (ALL) and the last 11 to the acute myeloid leukemia (AML) patients.

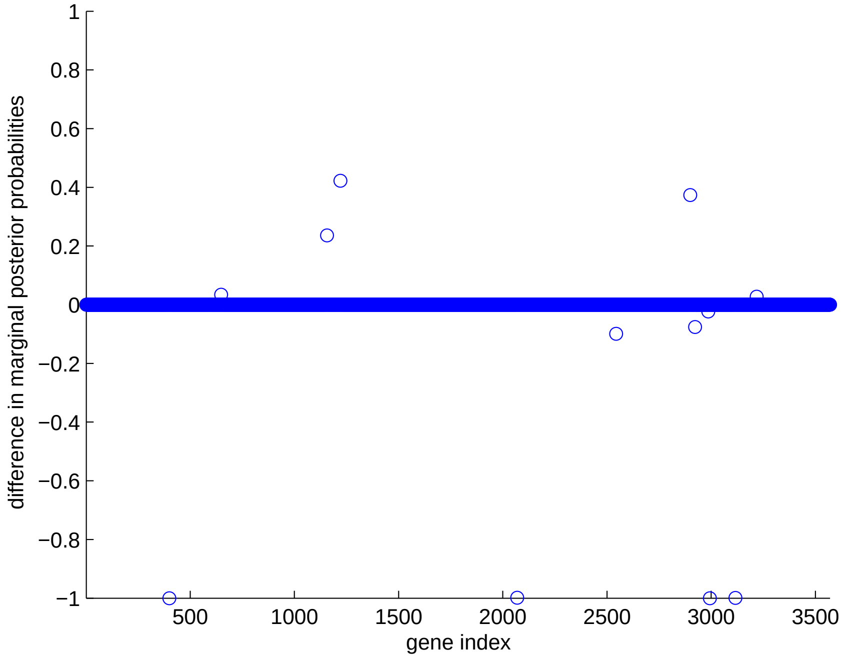


Figure 11: Microarray data: Concordance of results across MCMC chains

I note that, except for patient 25, and to a lesser extent patients 2, 12 and 20, all pairs of observations among the ALL group have a high probability of being assigned to the same cluster. The AML group instead exhibits less homogeneity. Thus, all

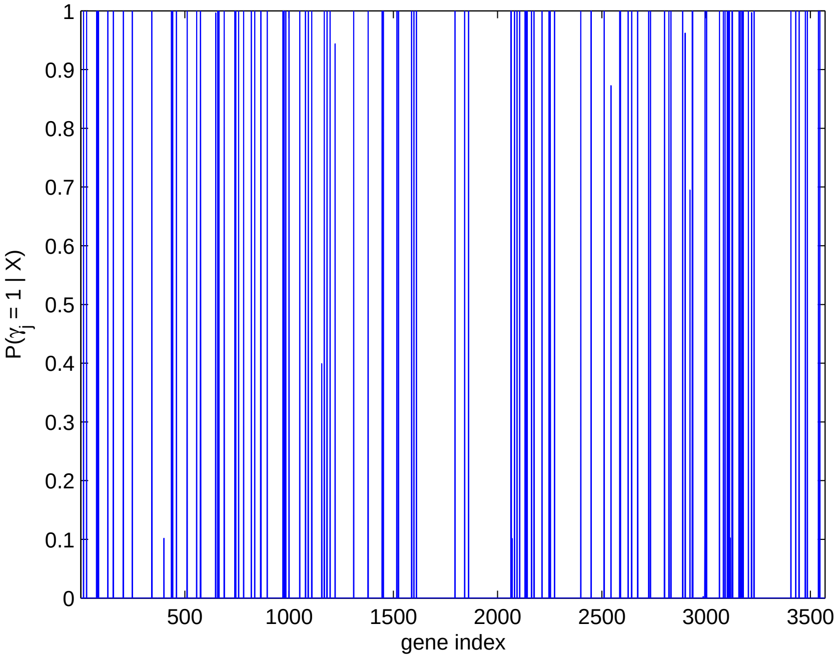


Figure 12: Microarray data: Marginal posterior probabilities $p(\gamma_j = 1|\boldsymbol{X})$

results indicate that I am able to successfully separate the ALL and AML patients and suggest that there are potential subgroups among the AML's.

For inference on the variable selection, I computed the marginal posterior probabilities of the γ_j's. Figure 11 displays the difference in marginal posterior probabilities for each gene across the two MCMC chains. I note that there is good concordance in the results despite the different starting points. This suggests that similar regions were visited by the two chains.

After pooling the output, I recomputed the marginal posterior probabilities, which are displayed in Figure 12. There were 116 genes with marginal posterior probabilities greater than 0.7. A heatmap of the selected genes is given in Figure 13, where the columns correspond to the leukemia patients and the rows represent the log gene expression levels. I note that these genes clearly discriminate the ALL (columns 1 to 27) and AML (columns 28 to 38) patients. The latter group is quite

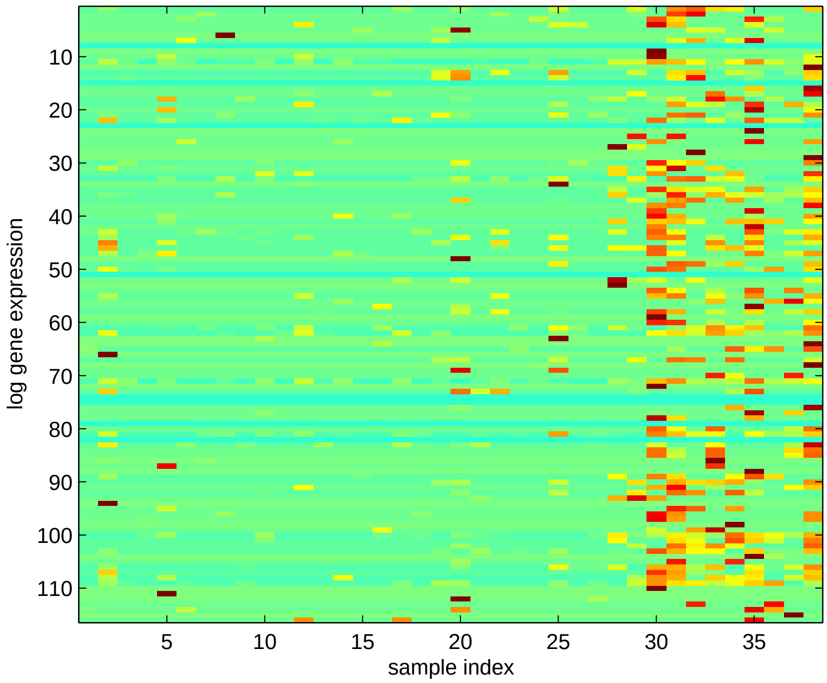


Figure 13: Heatmap of 116 genes selected by method

heterogeneous so that the result might provide the evidence of existence of subclasses in AML group.

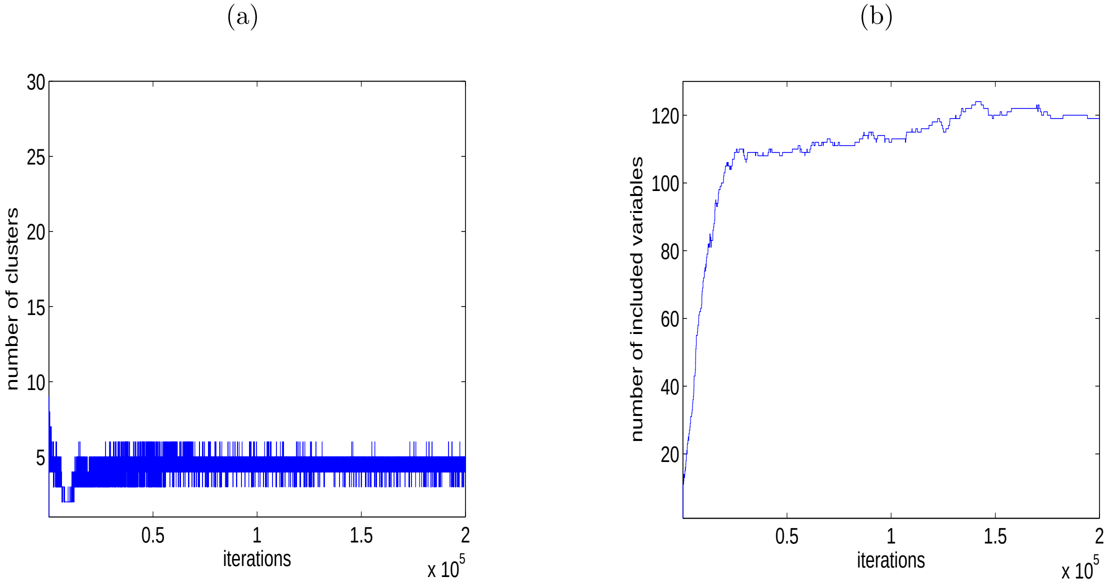
I also looked at the genes selected based on the $\widehat{\gamma}$ vector from equation (2.19). This set contained 120 genes that included all the 116 selected with the marginal inference. A large number of the genes identified by our method to discriminate the observations into the different subgroups are known to be implicated with the differentiation or progression of leukemia cells. I report some of the selected genes in Table 1.

Table 1: Some selected genes with known association to leukemia

Gene	Description
AR	The amphiregulin gene is localized in chromosomal region 4q13-4q21, a common breakpoint for ALL.
CA2	Expressed in most patients with leukemic blast cells.
CAV1	Hematological cells express caveolin-1 in certain states of cell activation and are believed to be a useful marker for adult T cell leukemia diagnosis.
CD1d&CD1c antigens	Significantly down-regulated in B cell chronic lymphocytic leukemia cells.
CD14 antigen	Maps to a region of chromosome 5 that contains a cluster of genes encoding several myeloid-specific growth factors and frequently deleted in certain myeloid leukemias.
CLC	Believed to be associated with myeloid cell differentiation into specific lineage leukemias and found to be significantly down-regulated in AML patients with high white blood cell count.
Cystatin A	Cystein protease inhibitor that induces apoptosis of leukemia cells.
ELA2	Elastase 2 cleaves the fusion protein generated by the translocation associated with promyelocytic leukemia.
ID4	Putative tumor suppressor silenced by promoter methylation in the majority of human leukemias.
IEX-1	Involved in modulation of apoptosis and highly expressed in acute promyelocytic leukemia cell lines.

Table 1: Continued

Gene	Description
IFN-alpha	Induces expression of myeloblastin and a specific T-cell response in chronic myeloid leukemia.
IL6 & IL8	These cytokines are elevated in activated T-cells in large granular lymphocytic leukemia.
LTF	Lactoferrin can transactivate Human T cell leukemia virus type I, which causes adult T cell leukemia and lymphoma.
MNDA	Correlated with myeloid and monocytic differentiation of acute leukemia, and expressed in M3 type leukemia but absent in ALL.
MT1G	The metalloprothionein gene cluster is mapped to 16p22, a breakpoint found in a subgroup of patients with AML.
PF4	Useful marker for categorizing megakaryocytic leukemic cells.
PRAME	Expressed in acute leukemia samples, with highest association in AML tumors carrying t(8;21) or t(15;17) chromosomal abnormalities that have a relatively favorable prognosis.
PTAFR	Increased level in eosinophilic leukemia cell line.
THBS1	Methylation of THBS1 is associated with the absence of the Philadelphia chromosome and a favorable prognosis for ALL patients.
TRAIL	Induces apostolic cell death in most chronic myelogenous and acute leukemia-derived Ph1-positive cell lines.
Tyrosine phosphatase 1	Primarily expressed in hematopoietic cells, it functions as an antagonist to the growth-promoting and oncogenic potentials of tyrosine kinase. It has been proposed as a candidate tumor suppressor gene in leukemia, lymphoma and other cancers.


(a)

(b)

Figure 14: Microarray data: Trace plots with $\alpha = 1$ for (a) Chain 2: number of clusters; (b) Chain 2: number of discriminating variables.

I repeated the analysis with $\alpha = 1$. Figure 14 shows trace plots of the number of clusters and the number of discriminating variables for one of the chains. It visited models with 3 to 6 clusters and around 120 discriminating variables. I note again a slightly slower mixing for smaller values of α, with the chain reaching 120 variables only at around iteration 140,000 (see Figures 9(b) and 9(d) for comparison with a larger α value). The posterior sample allocations were given by:

$$\text{using (2.17) } \widehat{\boldsymbol{c}} = (\underbrace{1, \ldots, 1, 1, 1, \ldots, 1, 1, 1, 1}_{ALL}, \underbrace{2, 2, 4, 3, 3, 2, 3, 6, 2, 2, 5}_{AML})$$

$$\text{using (2.18) } \widehat{\boldsymbol{c}} = (\underbrace{1, \ldots, 1, 2, 1, \ldots, 1, 2, 1, 1}_{ALL}, \underbrace{2, 2, 4, 5, 3, 2, 3, 6, 2, 2, 5}_{AML})$$

The ALL and AML samples are successfully separated. Samples 12 and 25 from the ALL class appear to be closer to some of the observations among the AML group. Again, I note more heterogeneity among the latter suggesting potential AML subtypes. The posterior inference on the variable selection identified 100 genes based on marginal posterior probabilities greater than 0.7, and 112 genes based on the $\widehat{\boldsymbol{\gamma}}$

vector with highest joint posterior probability. These were all included in the set of discriminating genes identified above.

3.3 Conclusion

I have applied the method described in chapter II to a simulated data and a DNA microarray data. The results from both applications were quite promising for discovering the correct cluster structures and for identifying discriminating genes. In the simulation study, the method uncovered the 4 true clusters with 17 selected discriminating variables. According to the result of the analysis on leukemia cancer DNA microarray data which consist of ALL and AML patients, most ALL patients have been successfully discovered. The result provided the evidence of existence of cancer subtypes among AML patients. From the literature, in fact, there exist 8 subtypes in AML, each of which has special and clinical laboratory features. About 120 genes selected by the method are known to be associated with progression or differentiation of leukemia cancer. As can be seen the result of application to leukemia DNA microarray data, I expect the method to be very useful to discover new subtypes of cancers. Eventually, this effort will contribute to improve the target-specific treatments for the patients with unrevealed cancer subtypes.

CHAPTER IV

CASE STUDY: WAVELET-BASED BAYESIAN CLUSTER ANALYSIS OF TIDAL VOLUME CURVES

4.1 Introduction

With advanced technology, high-dimensional data are commonly generated from a variety of research fields. One special type is high-dimensional functional data measured over a certain period of time. Functional data comprise functions as data for each subject. For example, in the bioinformatics area, temporal gene expression profiles discretely measured during cell cycle time provide a paramount information of gene function. In order to acquire knowledge of different biological funtions/pathways of genes, researchers are attentive to classify those gene expression profiles into small subgroups with different functional characteristics. In the analysis on this type of data, they consider an expression profile curve as a datum for each gene. Like this, the basic idea of functional data analysis (FDA) is that one should think of the observed data functions as single entities, rather than merely a sequence of individual observations. In practice, functional data are usually observed and recorded discretely even though there exist, in fact, countably infinite number of measurements in a latent true function.

One of the first steps in FDA is a data representation (in functional form) using interpolation and smoothing (Ramsay and Silverman 1997). If the discretely observed data are assumed to have no error, then the converting process is called *interpolation.* On the other hand, if they have some observational error that needs to be eliminated, then the converting procedure from the discrete data to functions may require *smoothing.* In this case, the measured data $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{ip})$ can be represented as

$$y_{ij} = f(t_j) + \epsilon_j, \quad j = 1, \cdots, p$$

where t_j is j-th time point and ϵ_j is the observational error at t_j. There are several techniques to represent functional data $\boldsymbol{y}_i$ as smooth functions. One of the most popular smoothing procedures involves representing the function by a linear combination of K known basis functions i.e.

$$f(t) = \sum_{k=1}^{K} c_k \varphi_k(t),$$

where c_k is a coefficient and φ_k is a basis function. There are many types of basis functions such as Fourier, polynomial, regression spline and wavelet bases. Among these many basis functions, how to choose a good basis function depends on how good the approximation is with relatively small number of basis functions and on the degree of information which coefficients provide about the data. In this chapter, I adopt wavelet bases since the data I will analyze are measured in time domain and look non-stationary or irregular. The first definition of wavelets can be attributed to Morlet, Arens, Fourgeau and Giard (1982). Currently, the term *wavelet* is usually associated with a wavelet function such that the translations and dilations of the function constitute an orthonormal basis. In other words, wavelets are families of orthonormal basis functions that can be used to parsimoniously represent other functions. For example, in $L^2(\mathbb{R})$, an orthogonal wavelet basis is obtained by dilating and translating a *mother wavelet* ψ as $\psi_{jk} = 2^{j/2}\psi(2^j x - k)$ with j, k integers. A function f can then be represented by the wavelet series $f(x) = \sum_{j,k} d_{jk}\psi_{jk}(x)$, with wavelet coefficients $d_{ik} = \int f(x)\psi_{jk}(x)dx$ describing features of the function f at the spatial locations indexed by k and scales indexed by j (see Appendix B for more information about wavelet transformations and wavelet shrinkage). Here I apply discrete wavelet transformations (DWTs) to project discrete data from the time domain to the wavelet domain.

In the next section, I introduce wavelet-based Bayesian clustering method based on the variable selection in cluster analysis described in the previous chapter. I then apply my method to tidal (or breath) volume measured on subjects under 3 different chemical injections that may induce panic attack. The research investigators want to examine the overall data for evidence of separation of the outcomes into groups. The cluster analysis does not use any knowledge about subjects belonging to different experimental conditions. This analysis was requested by a Data Monitoring Committee (DMC) as a preliminary step to provide an evidence that a certain chemical is an inducer of panic attacks by isolating the group experienced this chemical from the other groups.

4.2 Wavelet-based Bayesian clustering of functional Data

I extend a Bayesian method described in chapter II to high-dimensional functional data. The approach performs cluster analysis on the samples and selection of the relevant variables that discriminate observations. The functional adaption I exploit here results in a method that simultaneously reveals cluster structures among observations while identifying discriminating local features of the curves via the selection of the corresponding wavelet coefficients. My approach is model-based and uses Dirichlet process mixture priors that allow an infinite mixture of distributions to refine the observations revealing true cluster structures. I apply this clustering method to the wavelet decompositions of the data and add a variable selection mechanism to the model. The ability of wavelets to describe curve features at different levels of resolution gives us the option of selecting discriminating features at different time bandwidths.

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ be equal to a $n \times p$ matrix with each row representing a vector of observations of a function taken at p equally spaced points. Here, I assume p is power of 2. I apply a wavelet transform to each row of $\boldsymbol{Y}$. This results in a matrix $\boldsymbol{D} = \boldsymbol{Y}\boldsymbol{W}'$ of wavelet coefficients. I slightly abuse the notation here by assuming that $\boldsymbol{D}$ contains wavelet coefficients not affected by noise in the data. In other words, the matrix contains all scaling coefficients and survived wavelet coefficients after wavelet shrinkage (or thresholding) procedure. I model $\boldsymbol{D}$ as n independent observations $\boldsymbol{d}_i$ arising from a mixture of distribution $F(\boldsymbol{\theta}_i)$. Now I have the same models as I described in chapter II. In the previous chapter, cluster analysis has been conducted in data domain, but here all analysis would be done in the wavelet domain. A cluster analysis is completed by a Dirichlet process mixture model set up, where the creation and deletion of clusters is naturally taken care of in the process of updating the sample allocations and avoids computationally intensive algorithm, such as reversible jump MCMC. A DPM model provides the conditional prior probability of each configuration. With conjugate priors for model-specific parameters, the conditional posterior probabilities are easily calculated.

Variable selection, or feature selection is accomplished by introducing latent indicator vector $\boldsymbol{\gamma}$ with $\gamma_j = 1$ if j-th wavelet coefficient defines a mixture distribution and $\gamma_j = 0$ otherwise. Then matrix $\boldsymbol{D}$ is divided into two parts: one for helping clustering and the other with no information about clustering. Here, I denote by $\boldsymbol{D}_{(\gamma)}$ the set of coefficients that reveals cluster structures and by $\boldsymbol{D}_{(\gamma^c)}$ the remainder for unusable coefficients. Thus, $\boldsymbol{D}_{(\gamma)}$ has a mixture of distributions and $\boldsymbol{D}_{(\gamma^c)}$ favors one multivariate density across all observations. As I defined before, each row vector $\boldsymbol{d}_{i(\gamma)}$ of $\boldsymbol{D}_{(\gamma)}$ in cluster k is assumed to have a multivariate normal distribution with model parameters $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_{k(\gamma)}, \boldsymbol{\Sigma}_{k(\gamma)})$. All row $\boldsymbol{d}_{i(\gamma^c)}$s, $i = 1, \cdots, n$ have a single standard normal distribution $\mathcal{N}(\boldsymbol{\eta}_{(\gamma^c)}, \boldsymbol{\Omega}_{(\gamma^c)})$. Conjugate prior distributions have been adopted for mean vectors and covariance matrices.

In order to estimate cluster membership and selected coefficients, MCMC procedures are performed. For updating $\boldsymbol{\gamma}$, I adopt repeating a Metropolis algorithm several times. A split-merge algorithm for updating configuration vector for observations improves a mixing problem. Estimates of these two vectors are based on samples resulting through MCMC.

4.3 Case study: Panic attack

The functional adaptation describes here was motivated by a collaboration on an ongoing study that looks at high-dimensional, high-frequency measurements of breath tidal volume on a number of individuals, undergoing chemical interventions that may induce panic attacks. This is part of an interdisciplinary effort that involves investigators at the New York State Psychiatric Institute, at Columbia University. The overall goal is to create a model of the clinical panic attack in normal human subjects, as it occurs in individuals affected by panic disorder. The section consists of the following: data explanation, experimental design and goal, pre-processing of data, noise removal process by wavelets, and application of Bayesian variable selection via clustering. The case study is a preliminary step to see if there is a lactate effect inducing panic attacks. If cluster analysis confirms the separation of a group of subjects with the injection of lactate from the other observations, it will provide the evidence that supports the investigator's claim.

4.3.1 Data

Sodium lactate reliably produces panic attacks in patients with panic disorder (Liebowitz et al. 1985). Normal individuals who do not have panic disorder, rarely have such reactivity to lactate. A distinctive feature of lactate induced panic attack is a considerable increase in tidal volume (Goetz et al. 1993). Klein (1993) suggested that the spontaneous panic attack may be present due to a hypersensitive alarm system for detecting signals of impending suffocation, such as rising levels of CO2 or brain lactate. The endogenous opioid system is an important central regulator of respiratory drive. An exogenous opioid, such as morphine, blunts sensitivity to CO2 (Fleetham et al. 1980). Conversely, naloxone, an opioid receptor antagonist, increases the ventilatory response to hypercapnic hypoxia normal human controls (Akiyama et al. 1993). Naloxone pretreatment may make normal individuals who putatively have an intact opioid system, vulnerable to the marked anxiogenic and respiratory effects of lactate. In a pilot study, Sinha, Goetz, and Klein (2006) found that lactate after naloxone, administered to normal individuals, produced a marked increase in tidal volume that exceeded previous results from infusing only lactate. Surprisingly, lactate, despite producing a metabolic alkalosis, is a tidal volume stimulant, as has been shown in both normal humans and rats.

A randomized study with normal subjects was designed to test the investigators' hypothesis. Subjects, healthy normal male and female adult volunteers, not affected by either any psychiatric or significant medical illness, were randomized to three groups. They received either naloxone followed by lactate or saline followed by lactate or naloxone followed by saline. The hypothesis was that subjects receiving the naloxone-lactate sequence will have greater increases in tidal volume during the lactate phase than subjects in the other two groups. The naloxone-saline sequence should have lesser effects than the saline-lactate sequence. The randomization was unequal (3:3:1), with smallest number of subjects in the saline-lactate group, since prior experience with this sequence in normal subjects produced relatively minor increments in tidal volume. Establishing a lack of naloxone-saline effect was considered crucial. Respiratory and other physiological measurements were taken during the experiment, together with qualitative information measured via questionnaires and interviews. I report a cluster analysis performed on preliminary data from this ongoing study, when 50% of the planned sample size had completed the study. The cluster analysis do not use any knowledge about subjects belonging to different experimental conditions.

4.3.2 Experimental study

The experiment on each individual consisted of four phases:

(i) Phase I (baseline): Approximately 30 minutes. The subject has sensors and intravenous lines placed within 5 minutes while supine. This phase provides baseline measurements for each subject. Patency is maintained by slow saline drip, slowly increased to normal flow prior phase II. All infusion adjustments are made without the subject's knowledge. Personnel and subjects are blind to infusion contents. All randomized infusion sequences are set up in advance by the Research Pharmacist who maintains a secret subject listing.

(ii) Phase II (first infusion): Approximately 20 minutes. Subjects receive either naloxone over approximately the first 3 to 5 minutes, within the saline flow, or just stay on saline.

(iii) Phase III (second infusion): Approximately 20 minutes. Subjects who received naloxone at the first infusion are switched to either saline or lactate, and those that received only saline at the first infusion are switched to lactate. The infusion of the experimental component in the saline flow lasts approximately 20 minutes.

(iv) Phase IV (recovery): Approximately 120 minutes. The subject remains supine, with minimal saline flow. This period allows clinical observation as well as exploration of possible prolonged effects.

Thirty five subjects completed all phased of the experiment. The study is double-blinded since both investigators and volunteers do not have any knowledge of which chemicals were injected in both phase II and III. One subject showed a tidal volume (Vt) pattern that looked very different with respect to other subjects and was eliminated from the analysis. Recently, the true allocations of chemical (which chemical combination goes to whom) have been revealed after my analysis has been done. There were 15, 14 and 6 subjects in the three groups ”N+S” (Naloxone – Saline), ”N+L” (Naloxone – Lactate), and ”S+L” (Saline – Lactate), respectively. The removed one turned out to be in S+L group, and confirmed the subject react very peculiarly to injection comparing to others in the same group. The measuring and data recording device was the *lifeShirt* (Wilhelm, Roth and Sackner 2003), a garment recently developed with embedded inductive plethysmography sensors for continuous ambulatory monitoring of respiration and other physiological functions.

4.3.3 Pre-processing procedure

Each subject has a different Vt baseline. I therefore performed baseline adjustment by calculating the median Vt for each subject during phase I, which provides baseline measurements for each subject. Then I considered three ways to adjust for baseline effect: (a) subtracting the median from the Vt trace of each subject; (b) dividing the

Vt trace by the median; (c) taking log of (b). Results from the statistical analysis I performed did not show any particular sensitivity to these different procedures for baseline adjustment. Here I conduct the analysis with baseline-adjusted data using method (a).

Data are massive. During the experiment Vt measurements were automatically saved 50 times per second from the measurement device. I reduced dimension by considering traces obtained taking one every k-th data points. I examined plots of several reduced trace to make sure I was preserving important features of the data and decided on $k = 25$ as a safe choice. This gave me two measurements per second. In the analysis I considered data over a time window covering second infusion. For the second infusion, based on their previous experience with lactate infusions investigators do not expect a quick onset of effect during phase III. For the analysis I considered only the set of curves during second infusion because the hypothesized interactive effect of naloxone was expected during the subsequent infusion. Thus I used data measured over approximately 17 minutes before the end of the infusion. Figure 15 shows raw Vt with 51,200 data points in specified time window, covering second infusion. Figure 16 contains baseline adjusted Vt traces of all subjects with 2,048 data points, which are chosen at every 25-th in the whole time window.

4.3.4 Noise removal by wavelet methods

Processed traces looked very noisy as in Figure 16 with data I used for the analysis. Noise in the data can greatly affect clustering. Once one takes wavelet transformations of the data, the noise component can be easily removed. In my approach, I use a wavelet transformation also as a way to achieve a great reduction in dimension when fitting the clustering model which is performed in next section. The first step of the noise removal process is accomplished by applying a wavelet transform to the

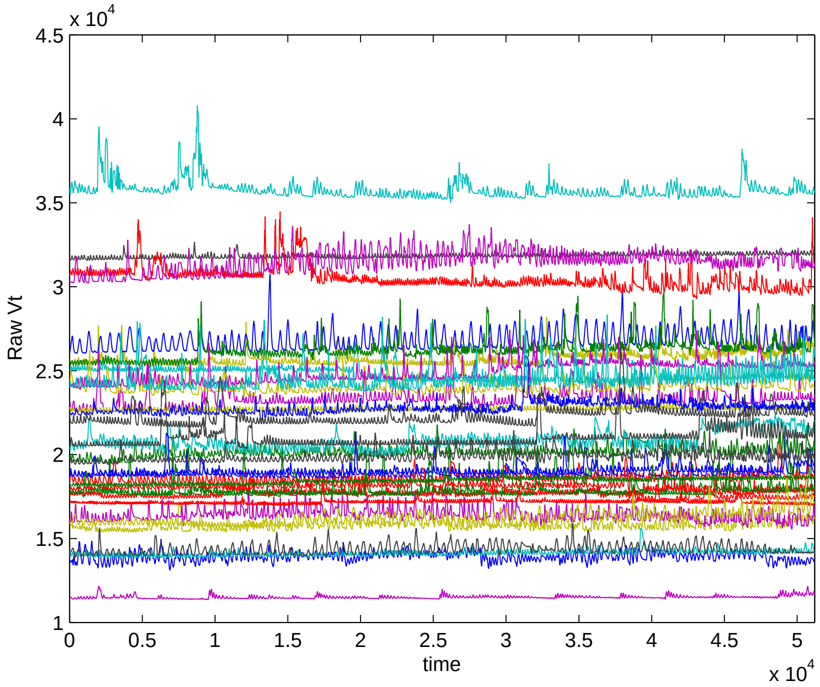


Figure 15: Raw data: 51,200 data points

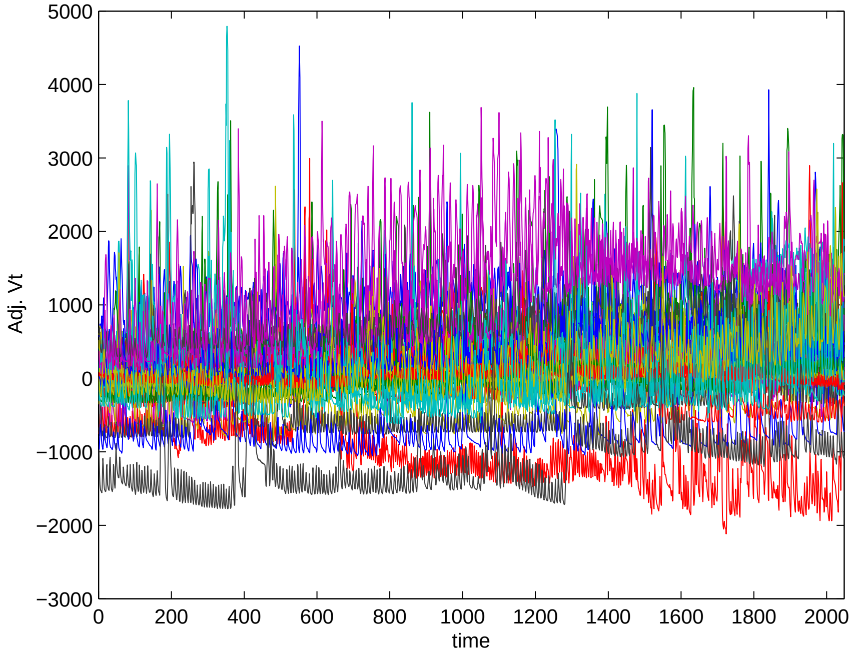


Figure 16: Baseline-adjusted Vt: 2,048 data points

data. Then one may do either mapping wavelet coefficients that fall below a chosen threshold to 0 (hard thresholidng) or shrinking the remaining coefficients toward 0 (soft thresholding). One can opt between universal or adaptive rules to choose the threshold. The former applies the same threshold, i.e. identical cut-off value for all wavelet coefficients, whereas the latter used a threshold that depends on the resolution level of the wavelet coefficients. An inverse wavelet transform is then applied to the thresholded coefficients leading to a smoothed estimate of the true signal.

In general, I noticed that the universal hard threshold removes lots of coefficients while the universal soft threshold tends to attenuate some of the distinctive features of the traces, such as the peaks. The adaptive soft thresholding approach, on the other hand, does a better job at preserving the peaks. As a result of this investigation, I chose to apply the Donoho and Johnstone *SureShrink* wavelet shrinkage with adaptive threshold. I used standard discrete wavelet transforms (DWT) with Daubechies wavelets with 4 vanishing moments. Wavelets with a higher number of vanishing moments have better regularity properties. On the other hand, the support of the wavelets increases with the regularity and boundary effects may arise in the DWT, so that a trade-off is often necessary.

By applying inverse discrete wavelet transforms (IDWT) to the wavelet coefficients that survive the thresholding step, I can reconstruct the Vt curves without noise. Figure 17 shows reconstructed curves for 3 randomly selected subjects, during the infusion. Note that how curve features show up more clearly in the reconstructed curves.

4.3.5 Bayesian cluster analysis

I performed the cluster analysis I described in the Section 3, to investigate whether the data would naturally separate into ”groups” or clusters. In the context, curves

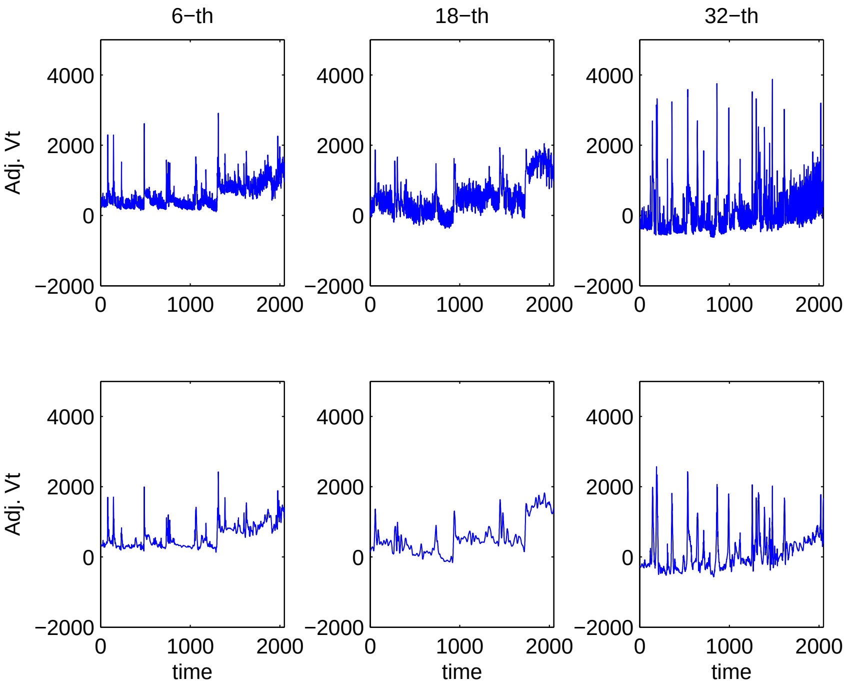


Figure 17: Reconstructed Vt with survived coefficients

belonging to the same group correspond to subjects that show similar Vt patterns during the time lag under consideration. For calibration purposes and in order to stabilize the variation in the data, I normalized the baseline adjusted Vt traces. This step is often crucial with multivariate statistical analyzes. Traces were normalized by subtracting the sample mean and dividing by the sample standard deviations.

I apply the non-parametric Bayesian approach for clustering to wavelet coefficients as described in previous Section. Clustering based on wavelet coefficients, rather than on curves, allows us to achieve a great reduction in dimensionality without losing critical information on the curves. A wavelet decomposition in fact can capture local features of curves with a small number of coefficients. Due to the properties of wavelets, wavelet coefficients describing local features will survive the procedure of noise removal. With my approach further dimension reduction is achieved via the selection mechanism built into the Bayesian procedure for clustering, where coefficients

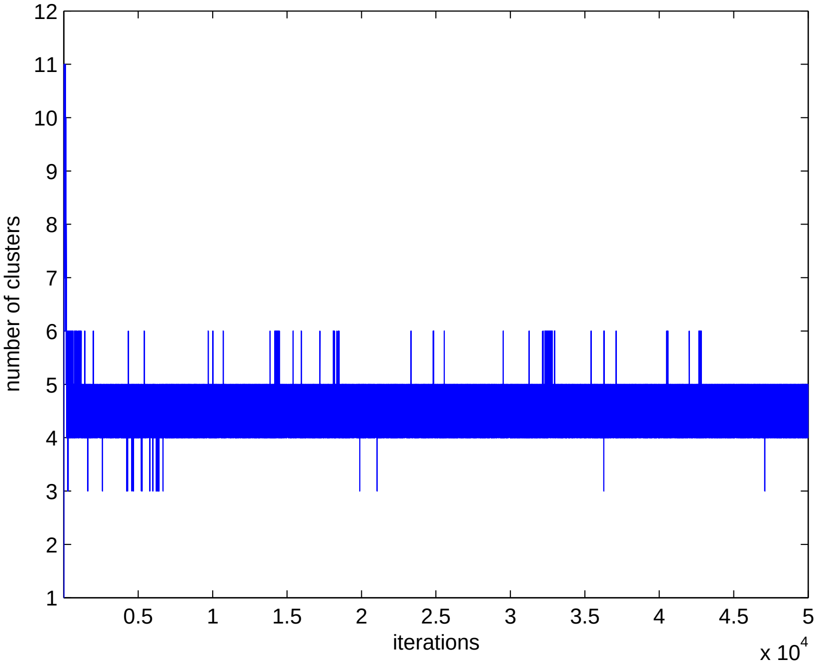


Figure 18: Trace plot of number of clusters

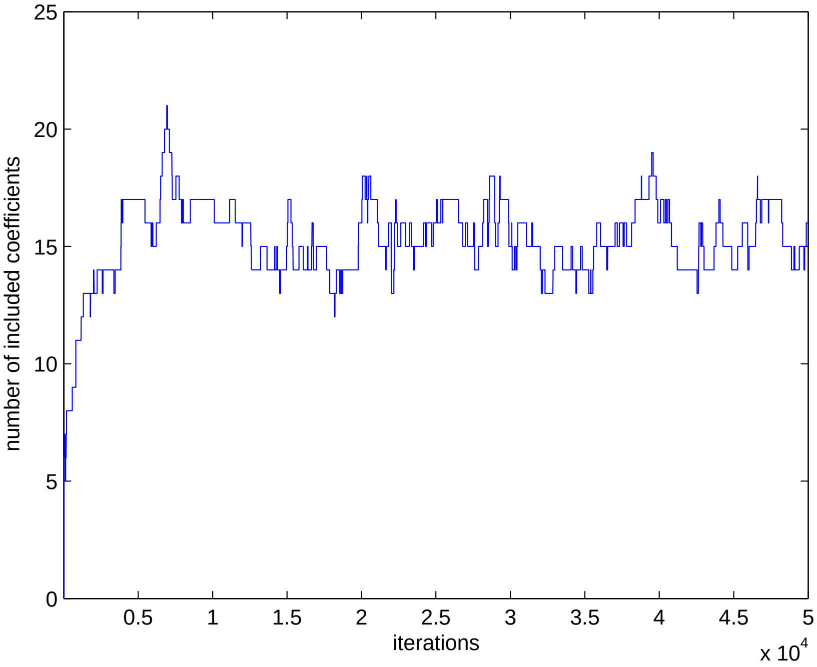


Figure 19: Trace plot of number of selected wavelet coefficients

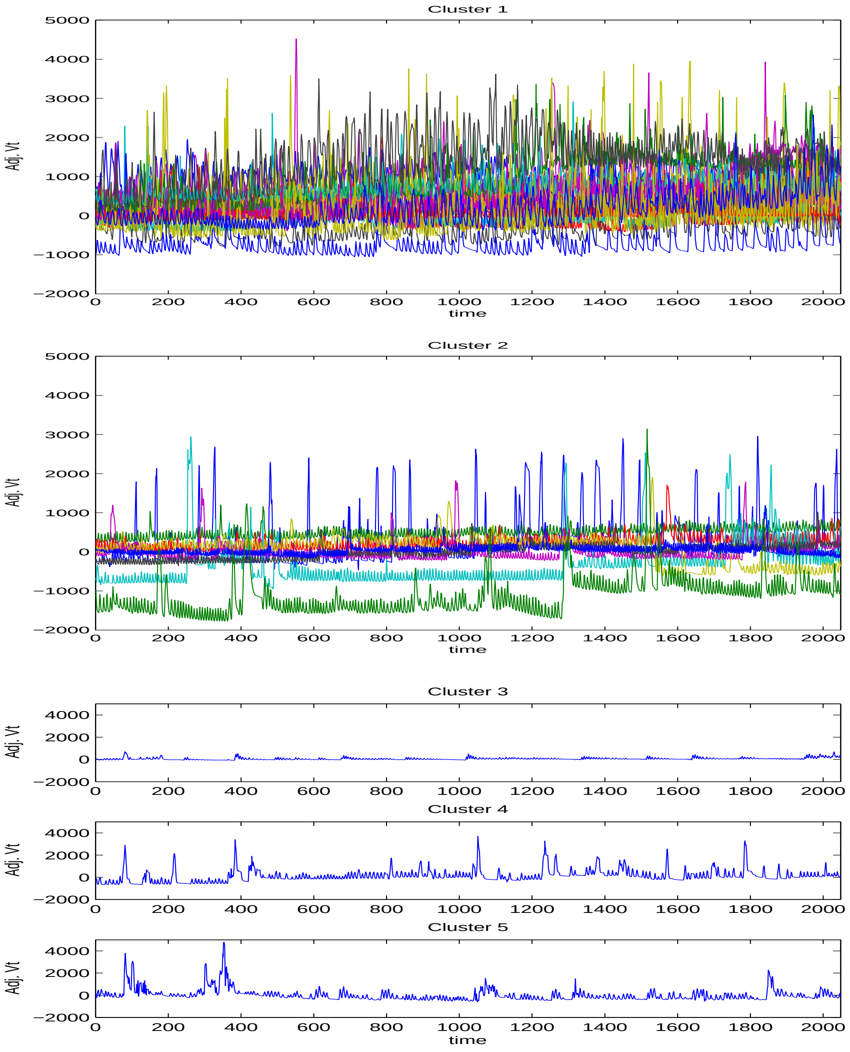


Figure 20: Cluster analysis during the 2nd infusion

that capture only the discrimination features of the curves are identified.

For the analysis here, I kept all coefficients that survived the thresholding procedure for at least 20% of the curves. This left approximately 500 wavelet coefficients for the analysis. Under the same criteria described in Chapter II, I set $\alpha = 34$, $h_0 = 100$, $h_1 = 100$, $\kappa_1 = 0.01$ and $b = 0.003$. I used 50,000 iterations with 20,000 burn-ins. Figure 18 and Figure 19 show plots of the MCMC traces for the number of clusters and the number of the included wavelet coefficients, respectively. The majority of the models visited during the MCMC had between 13 and 18 wavelet coefficients included. Most of the sampled cluster allocations had 4 to 5 clusters. As can be seen in figures, both samplers stabilize after 10,000 iterations.

In order to estimate both cluster membership and selected variables based on the MCMC output, I use the MAP method for estimating the configuration and marginal posterior probabilities for selecting variables. The MAP estimate for the cluster allocations resulted in one large cluster with 22 subjects, one smaller cluster with 9 subjects and 3 one-subject clusters. Figure 20 shows the baseline adjusted 2nd infusion traces plotted by cluster allocation. The large cluster with 22 subjects is depicted in the upper plot, the smaller cluster with 9 subjects in the middle plot and the 3 one-subject clusters in the lower plots. The results of this analysis suggest a clear separation of the curves. In particular, curves in the smaller clusters appear to have more irregular features, such as peaks, than those belonging to the bigger cluster.

Figure 21 shows the marginal posterior probabilities of γ_j. In order to gain insights on the differences in features among curves belonging to the different clusters, I looked at the scales of the selected wavelet coefficients. As previously described, coefficients at coarser levels of the wavelet transform describe features at lower frequency ranges and larger time periods. In this study, 15 wavelet coefficients were included in the "best" model, i.e., the subset of coefficients with highest posterior probability among those visited during the MCMC. The same subset was found by inspecting coefficients with largest marginal posterior probability of inclusion. Among these 15 selected coefficients, the majority belonged to the intermediate levels of the transform, in particular to the two levels that capture changes in the data over time periods of lengths of approximately $8\triangle t$ and $16\triangle t$, respectively, and frequency intervals $[\frac{1}{32\triangle t}, \frac{1}{16\triangle t}]$ and $[\frac{1}{64\triangle t}, \frac{1}{32\triangle t}]$, respectively, with $\triangle t = 0.5$sec. Discriminating

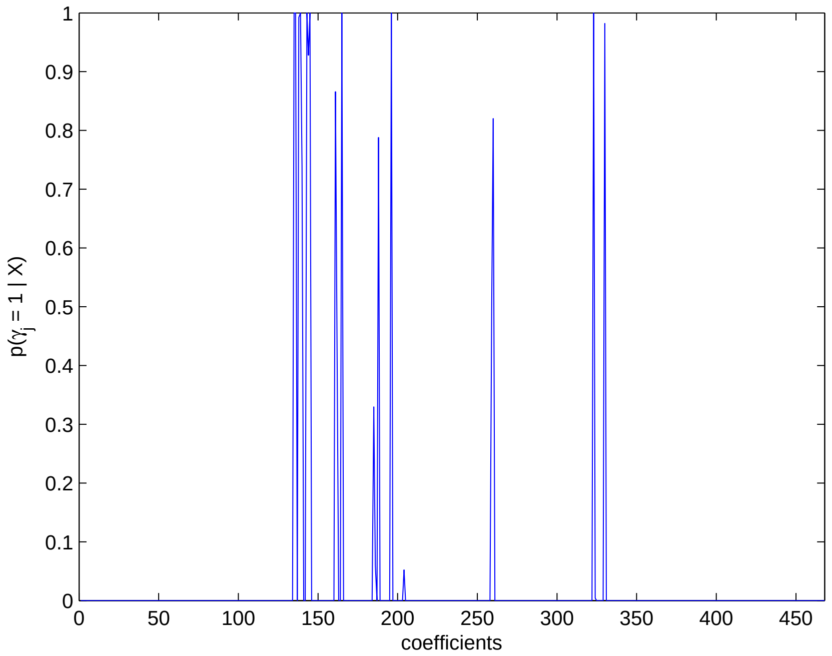


Figure 21: Marginal posterior probability of $\gamma_j = 1$

features of the curves are therefore local features spanning over relatively short time periods, such as the peaks of the smaller clusters.

I also looked at whether the clustering analysis showed any particular structure related to the interventions received by the subjects. Figure 22 contains group mean curves of cluster 1 with 22 subjects and cluster 2 of 9 samples. When comparing the cluster allocations with the group label information I noticed that 85% of the subjects receiving lactate during the second infusion belonged to the large cluster (cluster 1), and the 60% of the subjects on saline during the second infusion belonged to the smaller clusters. My results therefore supported a separation between the lactate and non-lactate groups during the second infusion.

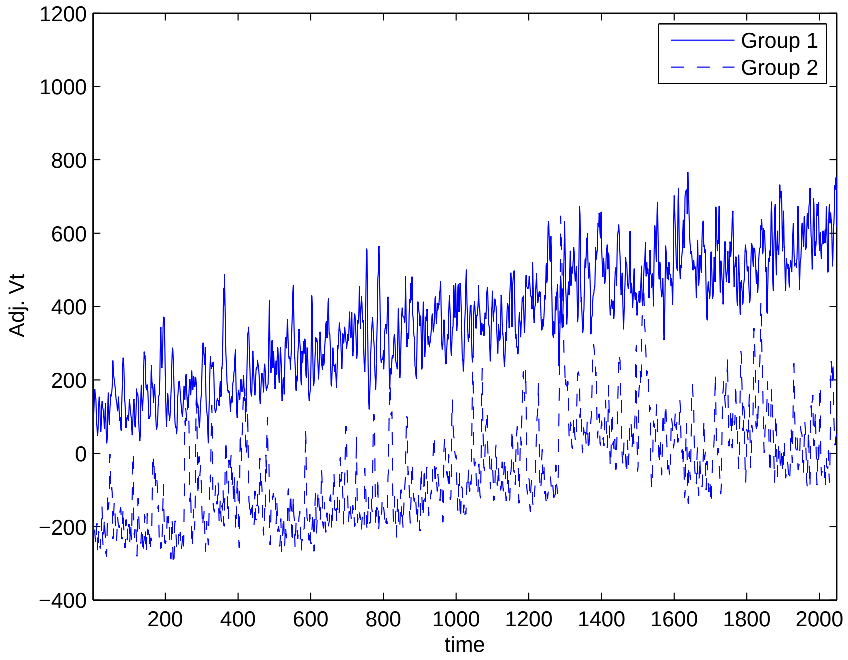


Figure 22: Group mean curves based on cluster allocations

4.4 Discussion

I have extended the variable selection in clustering via Dirichlet process mixture (DPM) models to high-dimensional functional data. Due to high-dimensionality and noise on functional data, I have applied a wavelet transformation (wavelet shrinkage) on the data to reduce dimensions and to remove noise. Then the model-based clustering method via DPM model is applied to distinguish individual curves into small subgroups with a few informative wavelet coefficients.

I specifically employed the method to assist investigators in a study of tidal volume traces during induced panic attacks. The study planned for an unblinded overall (not by treatment group) interim analysis of the efficacy data when 50% of the study samples had completed the interventions. The Data Monitoring Committee (DMC) wanted to examine the overall data for evidence of separation of the outcomes into groups. For the analysis I used a Bayesian procedure for model-based clustering that I adapted to functional data. The clustering analysis did not use any knowledge about subjects belonging to different experimental conditions. Classical clustering methods, such as linkage methods and k-means, highly depend on the distances among observations and often require the number of clusters to be pre-specified. I transformed curves into wavelet decompositions and performed cluster analysis based on wavelet coefficients. This allowed to remove noise from the data and to achieve a great reduction in dimensionality. My approach to clustering was model-based and used Dirichlet process mixture priors that allow an infinite mixture of distributions to refine observations to discover the correct cluster structures.

Results from the cluster analysis did suggest a separation of the subjects into groups during the time of the second infusion. In particular, I found one large cluster containing 22 subjects, a smaller cluster containing 9 subjects and 3 single-subject clusters. With respect to the actual treatment labels, the clustering analysis successfully isolated the lactate subjects from the non-lactate.

CHAPTER V

CONCLUSION

Thanks to advanced technology, high-dimensional data have arisen from diverse research fields, and especially in bioinformatics. Such data are characterized by a few observations and thousands of variables. Many classical statistical methods have been applied to answer various questions originating from high-dimensional data, such as class prediction for new observations and estimation of cluster structures. Such methods are often not feasible with high-dimensional data and also lead to incorrect estimations, due to too many irrelevant variables and redundant information. In cancer genomics, the main research interest of this dissertation, identification of discriminating genes of cancer subtypes is a crucial problem since potential subtypes of a cancer might respond differently to current target therapies. Tumors with similar histopathological appearance can follow significantly different courses and show different response to therapy. Thus it is essential to develop clustering methods to discover subclasses using a few informative variables. In this dissertation, I have developed a method that allows both variable selection and clustering to evolve simultaneously in search for the discriminating variables and the correct cluster structures. The methods I developed allow the same subset of variables across observations to discriminate the different clusters. Dirichlet process mixture (DPM) models allow an infinite mixture on the observations in order to find the correct cluster structures. This clustering method has several advantages: (i) there is no need to define a number of clusters before defining a model or to impose an explicit prior on the number of clusters in Bayesian framework; (ii) there is no need for computationally intensive algorithms such as reversible jump MCMC in order to update the cluster allocations.

To do variable selection, I have introduced a latent vector and adopted a stochastic search method to identify discriminating variables. I have evaluated my method on simulated data, leukemia cancer DNA microarray data and tidal volume functional data. The results from different application studies have confirmed that the methods are very useful. With high-dimensional data, I have expanded the developed approach by projecting time-functional data into wavelet domain.

Several extensions are possible for future work. First, there is a correlation problem I need to handle. In writing a likelihood, I assume the independence between discriminating and non-discriminating variables for computation convenience. In microarray study, it is true that some genes are highly correlated with other genes in their biological functional aspects. In other words, some genes which are identified as discriminating ones may be associated with other genes which are found to be non-discriminating. In this case, those discriminating genes may not be selected in the model due to high correlation with non-discriminating genes, or non-discriminating genes would be selected for the same reason. Currently it is not possible to include a correlation structure between these two types of variables in a model. But it is essential to consider the correlation in a model since there may be many discriminating genes which are correlated with ones in non-discriminating part or vice versa. In the future I expect that the method will be more improved to identify the most informative variables to reveal correct cluster structures. Another approach to handle this correlation problem is to employ correlation structure in a prior distribution of variable selection vector, $\boldsymbol{\gamma}$. In my current work, I adopt an independent Bernoulli distribution on each γ_j. For the same reason described above, this also needs improvement.

Next, combining two different types of methods in a model always slows and hinders the convergence of MCMC samplers. In usual, inclusion of variable selection in clustering make algorithm slow and inefficient. In my current approach, a Metropolis step is repeated several times in variable selection steps to allow samplers to stabilize for a given sample allocation. I adopt a Split-Merge algorithm through Metropolis-Hastings steps to solve this problem. In order to solve mixing problem more, I also add a tempering scheme to a Split-Merge. Exploiting these two or more algorithm may help to solve slow convergence and mixing problem. But for example, it is not easy to determine what distribution samplers are drawn from in parallel tempering case. Thus, it is necessary to develop a new algorithmic MCMC method for variable selection and cluster analysis to save computational time and improve mixing in mixture models.

Finally, it might be more accurate to find a way to incorporate other available covariates with target clustering data. For example, gene expressions measured on lung cancer tissue are highly heterogeneous. Thus, each observation tends to create its own clusters based on gene expression. Researchers in cancer studies have other informative variables rather than gene expression and want to take into account these variables in cluster analysis. If there is evidence between gene expressions and those variables, it is possible to set up a model for those variables in a linear relationship.

For my future study, I am currently working on varying-feature selection in clustering method. Varying-feature selection method has been widely studied in image analysis. The meaning of "varying-feature" is that there are different sets of features to distinguish different images. For example, when discriminating images of "clouds" and "lions" in a picture, one should use different partial features of each image. Sometimes it is more realistic to assume there are different features or variables to discriminate observations into different clusters. It is not just restricted to image analysis. For instance, there are 3 groups in the data, say Group 1, Group 2 and Group 3. There is a set of variables to discriminate Group 1 and Group 2, and another set of variables, which may or may not have some overlapping variables with the former variable group, to distinguish between Group 1 and Group 3. It would be the same for the other group comparisons. For example, in bioinformatics problem, it is in high demand to identify regulatory *cis*-elements (or motifs or Transcription Factor Binding Site (THBS)) to separate gene expressions into different group. But it may be ideal to spot the different sites for different functional groups of gene expression profiles. I am currently working to implement a varying feature selection method in classification/clustering on regression model set up.

Recently, the study on cluster analysis and/or variable selection method in high-dimensional data has become a very challenging area in statistics/bioinformatics. A Bayesian clustering approach via Dirichlet process mixture models has been one of the most promising methods to do cluster analysis. But there are still many mysterious doors to be opened to correctly discover cluster structures in high-dimensional data, and to identify discriminating variables.

REFERENCES

Akiyama, Y., Nishimura, M., Kobayashi, S., Yoshioka, A., Yamamoto, M., Miyamoto, K., and Kawakami, Y. (1993). "Effects of naloxone on the sensation of dyspnea during acute respiratory stress in normal adults." *Journal of Applied Physiology*, 74, 590–595.

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88, 669–679.

Antoniadis, A., Bigot, J., and Sapatinas, T. (2001). "Wavelet estimators in nonparametric regression: A comparative simulation study." *Journal of Statistical Software*, 6, 1–83.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2, 1152–1174.

Banfield, J. D. and Raftery, A. E. (1993). "Model-based Gaussian and non-Gaussian clustering." *Biometrics*, 49, 803–821.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1, 353–355.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*. Oxford: Clarendon.

Brown, P. J., Fearn, T., and Vannucci, M. (2001). "Bayesian wavelet regression on curves with application to a spectroscopic calibration problem." *Journal of American Statistical Association*, 96, 398–408.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). “Multivariate Bayesian variable selection and prediction.” *Journal of the Royal Statistical Society,* Ser. B, 60, 627–641.

Dahl, D. B. (2004). “Conjugate Dirichlet process mixture models: Efficient sampling, gene expression, and clustering.” PhD dissertation, University of Wisconsin, Department of Statistics.

——— (2006). “Model-based clustering for expression data via Dirichlet process mixture model.” in *Bayesian Inference for Gene Expression and Proteomics*, eds. K.-A. Do, P. Müller, and M. Vannucci. New York, NY: Cambridge University Press.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Number 61 in CBMS-NSF Series in Applied Mathematics. Philadelphia: SIAM.

Donoho, D. L. and Johnstone, I. M. (1994). “Ideal spatial adaption via wavelet shrinkage.” *Biometrika*, 82, 425–455.

——— (1998). “Minimax estimation via wavelet shrinkage.” *The Annals of Statistics*, 26, 879–921.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). “Comparison of discrimination methods for the classification of tumors using gene expression data.” *Journal of American Statistical Association*, 97, 77–87.

Escobar, M. D. (1994). “Estimating normal means with a Dirichlet process prior.” *Journal of American Statistical Association*, 89, 268–277.

Escobar, M. D. and West, M. (1995). “Bayesian density estimation and inference using mixtures.” *Journal of American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973). “A Bayesian analysis of some nonparametric problems.” *The Annals of Statistics*, 1, 209–230.

——— (1983). “Bayesian density estimation by mixtures of normal distributions.” in *Recent Advances in Statistics*, eds. H. Rizvi and J. Rustagi. New York: Academic Press.

Fleetham, J., Clarke, H., Dhingra, S., Chernick, V., and Anthiosen, N. (1980). “Endogenous opiates and chemical control of breathing in humans.” *American Review of Respiratory Disease*, 121, 1045–1049.

Friedman, J. H. and Meulman, J. J. (2004). “Clustering objects on subsets of attributes.” *Journal of the Royal Statistical Society,* Ser. B, 66, 815–849.

George, E. I. and McCulloch, R. E. (1993). “Variable selection via Gibbs sampling.” *Journal of American Statistical Association*, 88, 881–889.

——— (1997). “Approaches for Bayesian variable selection.” *Statistica Sinica*, 7, 339–373.

Geyer, C. J. (1991). “Markov Chain Monte Carlo maximum likelihood.” in *Computing Science and Statistics*, ed. E. M. Keramigas, 156–163. Fairfax: Interface Foundation.

Goetz, R., Klein, D., Gully, D., Kahn, J., Liebowitz, M., Fyer, A., and Gorman, J. (1993). “Panic attacks during placebo procedures in the laboratory: Physiology and symptomatology.” *Archives of General Psychiatry*, 50, 280–285.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and

Lander, E. S. (1999). “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.” *Science*, 286, 531–537.

Green, P. J. and Richardson, S. (2001). “Modeling heterogeneity with and without the Dirichlet process.” *Scandinavian Journal of Statistics*, 28, 355–375.

Hoff, P. D. (2006). “Model-based subspace clustering.” *Bayesian Analysis 1*, 321–344.

Ishwaran, H. and James, L. F. (2001). “Gibbs sampling methods for stick-breaking priors.” *Journal of American Statistical Association*, 96, 161–173.

Jain, S. and Neal, R. M. (2004). “A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.” *Journal of Computational and Graphical Statistics*, 13, 158–182.

Kass, R. E. and Wasserman, L. (1996). “The selection of prior distributions by formal rules.” *Journal of American Statistical Association*, 91, 1343–1370.

Klein, D. (1993). “False suffocation alarms, spontaneous panics and related conditions: An integrative hypothesis.” *Archives of General Psychiatry*, 50, 306–317.

Liebowitz, M., Gorman, J., Fyer, A., Levitt, M., Dillon, D., Levy, G., Appleby, I., Anderson, S., Palij, M., Davies, S., and Klein, D. (1985). “Lactate provocation of panic attacks: Clinical and behavioral findings.” *Archives of General Psychiatry*, 41, 764–770.

Liu, J. S., Zhang, J. L., Palumbo, M. L., and Lawrence, C. E. (2003). “Bayesian clustering with variable and transformation selections.” in *Bayesian Statistics*, eds. J. M. Bernardo, M. J. Bayari, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 249–275. New York, NY: Oxford University Press.

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics: Simulation and Computation*, 23, 727–741.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999). "Sequential importance sampling for nonparametric Bayes models: The next generation." *The Canadian Journal of Statistics*, 27, 251–267.

MacEachern, S. N. and Müller, P. (1998). "Estimating mixtures of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7, 223–238.

Madigan, D. and York, J. (1995). "Bayesian graphical models for discrete data." *International Statistical Review*, 63, 215–232.

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

Medvedovic, M. and Sivaganesan, S. (2002). "Bayesian infinite mixture model based clustering of gene expression profiles." *Bioinformatics*, 18, 1194–1206.

Miller, A. (1990). *Subset Selection in Regression*. London: Chapman & Hall.

Morlet, J., Arens, G., Fourgeau, E., and Giard, D. (1982). "Wave propagation and sampling theory." *Geophysics*, 47, 203–236.

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003). "Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis (with discussion)." *Journal of the American Statistical Association*, 98, 573–597.

Neal, R. M. (1992). "Bayesian Mixture Modeling." in *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, 1991*, eds.

C. Smith, G. J. Erickson, and P. O. Neudorfer, 197–211. Dordrecht, The Netherlands: Kluwer Academic Publishers.

——— (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9, 249–265.

Percival, D. B. and Walden, A. T. (1993). *Wavelet Methods for Time Series Analysis*. Cambridge, U.K.: Cambridge University Press.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society,* Ser. B, 59, 731–792.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4, 639–650.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage." *Biometrics*, 60, 812–819.

Sinha, S., Goetz, R., and Klein, D. (2006). "Physiological and behavioral effects of naloxone and lactate in normals with relevance to the pathophysiology of panic disorder." *Psychiatry Research*, in press.

Stephens, M. (2000a). "Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods." *The Annals of Statistics*, 28, 40–74.

——— (2000b). “Dealing with label switching in mixture models.” *Journal of the Royal Statistical Society,* Ser. B, 62, 795–809.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). “Bayesian variable selection in clustering high-dimensional data.” *Journal of American Statistical Association*, 100, 602–617.

Vannucci, M., Sha., N., and Brown, P. J. (2005). “NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection.” *Chemometrics and Intelligent Laboratory Systems*, 77, 139–148.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York, NY: Wiley.

Wasserman, L. (2000). “Asymptotic inference for mixture models using data-dependent priors.” *Journal of the Royal Statistical Society,* Ser. B, 62, 159–180.

Wilhelm, F. H., Roth, W. T., and Sackner, M. A. (2003). “The lifeShirt: An advanced system for ambulatory measurement of respiratory and cardiac function.” *Behavior Modification*, 27, 671–691.

APPENDIX A

The purpose of this appendix is to give details on the derivations of marginalized full conditionals (2.12) and $A(c_i)$ in (2.16) Chapter II. III .

1. Marginalized full conditional distribution in (2.12)

$$\begin{aligned}
f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{c}) &= 2^{-\frac{n(p-p_\gamma)}{2}}\pi^{-\frac{np}{2}}\prod_{k=1}^{K}\left[\boldsymbol{H}_{k(\gamma)}\cdot|\boldsymbol{Q}_{1(\gamma)}|^{\frac{\delta+p_\gamma-1}{2}}\cdot\left|\boldsymbol{Q}_{1(\gamma)}+\boldsymbol{S}_{k(\gamma)}\right|^{-\frac{n_k+\delta+p_\gamma-1}{2}}\right]\\
&\times \boldsymbol{H}_{0(\gamma^c)}\cdot\left[\boldsymbol{S}_{0(\gamma^c)}\right]^{-(a+n/2)},
\end{aligned}$$

$$\begin{aligned}
\text{where } \boldsymbol{H}_{k(\gamma)} &= (h_1n_k+1)^{-\frac{p_\gamma}{2}}\prod_{j=1}^{p_\gamma}\frac{\Gamma\left(\frac{n_k+\delta+p_\gamma-j}{2}\right)}{\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)},\\
\boldsymbol{H}_{0(\gamma^c)} &= (h_0n+1)^{-\frac{p-p_\gamma}{2}}b^{a(p-p_\gamma)}\prod_{j=1}^{p-p_\gamma}\frac{\Gamma\left(a+n/2\right)}{\Gamma(a)},\\
\boldsymbol{S}_{k(\gamma)} &= \sum_{i\in C_k}(\boldsymbol{x}_{i(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})(\boldsymbol{x}_{i(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})^T\\
&+ \frac{n_k}{h_1n_k+1}(\boldsymbol{\mu}_{0(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})(\boldsymbol{\mu}_{0(\gamma)}-\bar{\boldsymbol{x}}_{k(\gamma)})^T,\\
\boldsymbol{S}_{0(\gamma^c)} &= \prod_{j=1}^{p-p_\gamma}\left[b+\frac{1}{2}\left\{\sum_{i=1}^{n}(x_{ij(\gamma^c)}-\bar{x}_{j(\gamma^c)})^2+\frac{n}{h_0n+1}(\mu_{0j(\gamma^c)}-\bar{x}_{j(\gamma^c)})^2\right\}\right],
\end{aligned}$$

with $\bar{\boldsymbol{x}}_{k(\gamma)}$ the sample mean of cluster k, and $\bar{x}_{j(\gamma^c)}$ the j-th non-discriminating variable sample mean.

<Derivation>

$$\begin{aligned}
f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{c}) &= \int_{\sigma^2}\int_{\boldsymbol{\Sigma}}\int_{\boldsymbol{\eta}}\int_{\boldsymbol{\mu}} f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{\eta},\boldsymbol{\Sigma},\sigma^2)f(\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{\gamma})f(\boldsymbol{\Sigma}|\boldsymbol{\gamma})\\
&\times f(\boldsymbol{\eta}|\boldsymbol{\Omega},\boldsymbol{\gamma})f(\boldsymbol{\Omega}|\boldsymbol{\gamma})\ d\boldsymbol{\mu}d\boldsymbol{\eta}d\Sigma d\sigma^2\\
&= \int_{\sigma^2}\int_{\boldsymbol{\eta}} f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{\eta},\sigma^2)f(\boldsymbol{\eta}|\boldsymbol{\gamma},\sigma^2)\prod_{j=1}^{p-p_\gamma}f(\sigma^2)\ d\boldsymbol{\eta}d\sigma^2\\
&\times \int_{\boldsymbol{\Sigma}}\int_{\boldsymbol{\mu}} f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{\Sigma})f(\boldsymbol{\mu}|\boldsymbol{\gamma},\boldsymbol{\Sigma})f(\boldsymbol{\Sigma}|\boldsymbol{\gamma})\ d\boldsymbol{\mu}d\boldsymbol{\Sigma}
\end{aligned}$$

The former integration part above is non-discriminating part and the latter discriminating part. I assume the independence between these two types of variables. I will first calculate the integration related to non-discriminating variables.

$$
\begin{aligned}
& \int_{\sigma^2}\int_{\boldsymbol{\eta}} f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{\eta},\sigma^2) f(\boldsymbol{\eta}|\boldsymbol{\gamma},\sigma^2) \prod_{j=1}^{p-p_\gamma} f(\sigma^2)\; d\boldsymbol{\eta} d\sigma^2 \\
= \;& \prod_{i=1}^{n}\int_{\sigma^2}\int_{\boldsymbol{\eta}_{(\gamma^c)}} \left[(2\pi)^{-\frac{p-p_\gamma}{2}} |h_0\sigma^2 I|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2h_0\sigma^2}(\boldsymbol{x}_{i(\gamma^c)}-\boldsymbol{\eta}_{(\gamma^c)})^t(\boldsymbol{x}_{i(\gamma^c)}-\boldsymbol{\eta}_{(\gamma^c)})\right\}\right] \\
\times \;& (2\pi)^{-\frac{p-p_\gamma}{2}} |\sigma^2 I|^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\eta}_{(\gamma^c)}-\boldsymbol{\mu}_{0(\gamma^c)})^t(\boldsymbol{\eta}_{(\gamma^c)}-\boldsymbol{\mu}_{0(\gamma^c)})\} \\
\times \;& \prod_{j=1}^{p-p_\gamma}\left[\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp\{-\frac{b}{\sigma^2}\}\right]\; d\boldsymbol{\eta}_{(\gamma^c)} d\sigma^2 \\
= \;& (2\pi)^{-\frac{(n+1)(p-p_\gamma)}{2}} \left(\frac{b^a}{\Gamma(a)}\right)^{p-p_\gamma} h_0^{-\frac{n(p-p_\gamma)}{2}} \\
\times \;& \int_{\sigma^2}(\sigma^2)^{-\left(\frac{(p-p_\gamma)(n+1)}{2}+a+1\right)} \exp\left\{-\frac{(p-p_\gamma)b}{\sigma^2}-\frac{1}{2\sigma^2}\left(\boldsymbol{\mu}_{0(\gamma^c)}^t\boldsymbol{\mu}_{0(\gamma^c)}+\frac{1}{h_0}\sum_{i=1}^{n}\boldsymbol{x}_{i(\gamma^c)}^t\boldsymbol{x}_{i(\gamma^c)}\right)\right\} \\
\times \;& \int_{\boldsymbol{\eta}_{(\gamma^c)}} \exp\left[-\frac{1}{2h_0\sigma^2}\left\{(1+h_0)\boldsymbol{\eta}_{(\gamma^c)}^t\boldsymbol{\eta}_{(\gamma^c)}-2\left(\sum_{i=1}^{n}\boldsymbol{x}_{i(\gamma^c)}+h_0\boldsymbol{\mu}_{0(\gamma^c)}\right)^t\boldsymbol{\eta}_{(\gamma^c)}\right\}\right]\; d\boldsymbol{\eta}_{(\gamma^c)} d\sigma^2 \\
= \;& (2\pi)^{-\frac{n(p-p_\gamma)}{2}} \left(\frac{b^a}{\Gamma(a)}\right)^{p-p_\gamma} (h_0 n+1)^{-\frac{p-p_\gamma}{2}} \int_{\sigma^2}(\sigma^2)^{-(n+\frac{a}{2}+1)}\exp\left\{-\frac{1}{\sigma^2}\boldsymbol{S}_{0(\gamma^c)}\right\}\; d\sigma^2 \\
= \;& (2\pi)^{-\frac{n(p-p_\gamma)}{2}} \boldsymbol{H}_{0(\gamma^c)}\; [\boldsymbol{S}_{0(\gamma^c)}]^{-(n+\frac{a}{2})}
\end{aligned}
$$

Similarly, the integration of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is derived as follows:

$$
\begin{aligned}
& \int_{\boldsymbol{\Sigma}}\int_{\boldsymbol{\mu}} f(\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{\Sigma}) f(\boldsymbol{\mu}|\boldsymbol{\gamma},\boldsymbol{c},\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}|\boldsymbol{\gamma},\boldsymbol{c})\; d\boldsymbol{\mu} d\boldsymbol{\Sigma} \\
= \;& \prod_{k=1}^{K}\int_{\boldsymbol{\Sigma}_{k(\gamma)}}\int_{\boldsymbol{\mu}_{k(\gamma)}} (2\pi)^{-\frac{n_k p_\gamma}{2}} |h_1\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{n_k}{2}} \exp\left\{-\frac{1}{2h_1}\sum_{i\in C_k}(\boldsymbol{x}_{i(\gamma)}-\boldsymbol{\mu}_{k(\gamma)})^t\boldsymbol{\Sigma}_{k(\gamma)}^{-1}(\boldsymbol{x}_{i(\gamma)}-\boldsymbol{\mu}_{k(\gamma)})\right\} \\
\times \;& (2\pi)^{-\frac{p_\gamma}{2}} |\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_{k(\gamma)}-\boldsymbol{\mu}_0)^t\boldsymbol{\Sigma}_{k(\gamma)}^{-1}(\boldsymbol{\mu}_{k(\gamma)}-\boldsymbol{\mu}_{0(\gamma)})\right\} \\
\times \;& 2^{-\frac{(\delta+p_\gamma-1)p_\gamma}{2}}\pi^{-\frac{p_\gamma(1-p_\gamma)}{4}}\frac{1}{\prod_{j=1}^{p_\gamma}\Gamma(\frac{\delta+p_\gamma-1}{2})}|\boldsymbol{Q}_1|^{\frac{\delta+p-1}{2}}|\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{\delta+2p}{2}}\; d\boldsymbol{\mu}_{k(\gamma)} d\boldsymbol{\Sigma}_{k(\gamma)}
\end{aligned}
$$

$$
\begin{aligned}
& = \int_{\boldsymbol{\Sigma}}\int_{\boldsymbol{\mu}}\prod_{k=1}^{K}\left\{(2\pi)^{-\frac{(n_k+1)p_\gamma}{2}}h_1{}^{-\frac{p_\gamma}{2}}\frac{2^{-\frac{(\delta+p_\gamma-1)p_\gamma}{2}}\pi^{-\frac{p_\gamma(1-p_\gamma)}{4}}}{\prod_{j=1}^{p_\gamma}\Gamma(\frac{\delta+p_\gamma-1}{2})}|\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{n_k+1}{2}}\right\} \\
\times \quad & \exp\left\{-\frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}(\boldsymbol{x}_{i(\gamma)}-\boldsymbol{\mu}_{k(\gamma)})^t\boldsymbol{\Sigma}_{k(\gamma)}^{-1}(\boldsymbol{x}_{i(\gamma)}-\boldsymbol{\mu}_{k(\gamma)})\right\} \\
\times \quad & \exp\left\{-\frac{1}{2}\sum_{k=1}^{K}\left(\boldsymbol{\mu}_{k(\gamma)}-\frac{\sum_{i\in C_k}\boldsymbol{x}_{i(\gamma)}+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}}{n_k+h_1^{-1}}\right)^t(n_k+h_1^{-1})\boldsymbol{\Sigma}_{k(\gamma)}^{-1}\right. \\
& \left.\left(\boldsymbol{\mu}_{k(\gamma)}-\frac{\sum_{i\in C_k}\boldsymbol{x}_{i(\gamma)}+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}}{n_k+h_1^{-1}}\right)\right\} \\
\times \quad & \prod_{k=1}^{K}\left[|\boldsymbol{Q}_{1(\gamma)}|^{\frac{\delta+p_\gamma-1}{2}}|\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{\delta+2p_\gamma}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{k(\gamma)}^{-1}(\boldsymbol{Q}_{1(\gamma)}+\boldsymbol{S}_{k(\gamma)})\right)\right\}\right]\ d\boldsymbol{\mu}d\boldsymbol{\Sigma} \\
= \quad & \pi^{-\frac{np_\gamma}{2}}\prod_{k=1}^{K}\left\{(h_1n_k+1)^{-\frac{p_\gamma}{2}}|\boldsymbol{Q}_{1(\gamma)}|^{\frac{\delta+p_\gamma-1}{2}}\frac{2^{-\frac{(\delta+p_\gamma-1)p_\gamma}{2}}\pi^{-\frac{p_\gamma(1-p_\gamma)}{4}}}{\prod_{j=1}^{p_\gamma}\Gamma(\frac{\delta+p_\gamma-1}{2})}\right\} \\
& \times\int_{\boldsymbol{\Sigma}}\prod_{k=1}^{K}\left[|\boldsymbol{\Sigma}_{k(\gamma)}|^{-\frac{n_k+\delta+2p_\gamma}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{k(\gamma)}^{-1}(\boldsymbol{Q}_{1(\gamma)}+\boldsymbol{S}_{k(\gamma)})\right)\right\}\right]\ d\boldsymbol{\Sigma} \\
= \quad & \pi^{-\frac{np_\gamma}{2}}\prod_{k=1}^{K}\left\{(h_1n_k+1)^{-\frac{p_\gamma}{2}}\prod_{j=1}^{p_\gamma}\left(\frac{\Gamma(\frac{n_k+\delta+p_\gamma-1}{2})}{\Gamma(\frac{\delta+p_\gamma-1}{2})}\right)\right\} \\
& \times\prod_{k=1}^{K}\left\{|\boldsymbol{Q}_{1(\gamma)}|^{\frac{\delta+p_\gamma-1}{2}}|\boldsymbol{Q}_{1(\gamma)}+\boldsymbol{S}_{k(\gamma)}|^{-\frac{n_k+\delta+p_\gamma-1}{2}}\right\}.
\end{aligned}
$$

2. $A(c_i)$ **in (2.16)**

$$A(c_i) = \int F(\boldsymbol{x}_k; \boldsymbol{\phi}, \boldsymbol{\gamma}) dH_{-k,c_i}(\boldsymbol{\phi}, \boldsymbol{\gamma}) =$$

$$\pi^{-p_\gamma/2} \left(\frac{h_1 n_{c_i} + 1}{h_1 n_{-k,c_i} + 1} \right)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma\left(\frac{n_{c_i}+\delta+p_\gamma-j}{2}\right)}{\Gamma\left(\frac{n_{-k,c_i}+\delta+p_\gamma-j}{2}\right)}$$

$$\times \left|\boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{c_i(\gamma)}\right|^{-(n_{c_i}+\delta+p_\gamma-1)/2} \left|\boldsymbol{Q}_{1(\gamma)} + \boldsymbol{S}_{-k,c_i(\gamma)}\right|^{(n_{-k,c_i}+\delta+p_\gamma-1)/2},$$

$$\begin{aligned} \text{with } \boldsymbol{S}_{-k,c_i(\gamma)} &= \textstyle\sum_{j \neq k: c_j = c_i} \left(\boldsymbol{x}_{j(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)}\right) \left(\boldsymbol{x}_{j(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)}\right)^T \\ &+ \tfrac{n_{-k,c_i}}{h_1 n_{-k,c_i}+1} \left(\boldsymbol{\mu}_{0(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)}\right) \left(\boldsymbol{\mu}_{0(\gamma)} - \bar{\boldsymbol{x}}_{c_i(\gamma)}\right)^T \end{aligned}$$

and $\boldsymbol{S}_{c_i(\gamma)}$ is defined as in equation (2.12).

<Derivation>

$$\begin{aligned} \int F(\boldsymbol{x}_k); \boldsymbol{\phi}, \boldsymbol{\gamma}) dH_{-k,c_i}(\boldsymbol{\phi}, \boldsymbol{\gamma}) &= \int F(\boldsymbol{x}_k; \boldsymbol{\phi}, \boldsymbol{\gamma}) \frac{F(\boldsymbol{x}_{-k}; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})}{\int F(\boldsymbol{x}_{-k,c_i}; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})} \\ &= \frac{\int F(\boldsymbol{x}_{c_i}; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})}{\int F(\boldsymbol{x}_{-k,c_i}; \boldsymbol{\phi}, \boldsymbol{\gamma}) dG_0(\boldsymbol{\phi}, \boldsymbol{\gamma})}. \end{aligned}$$

The difference between top and bottom of the fraction above is whether data $\boldsymbol{x}$ used in calculation includes elements k in $\mathcal{C}$. The top part includes all observations with same configuration c_i, and the bottom excludes the elements in $\mathcal{C}$ with configuration c_i. The calculation is very obvious from the marginalized likelihood above.

APPENDIX B

The purpose of this appendix is to give the introduction of wavelets needed to understand Chapter IV .

1. Some background of Wavelets

It is well-known that a function f can be represented by a set of orthogonal basis functions. Wavelets are sets of orthonormal basis functions generated by dilation and translations of basic parent function: *scaling function* (also called father wavelet) ϕ which typically resembles a kernel function, and *mother wavelet* (also called wavelet function) ψ which explains oscillation. For better understanding for wavelets, I will start with the explanation of *multiresolution analysis* (MRA). A MRA is a sequence of closed subspaces V_n, $n \in \mathbb{Z}$ in $L^2(\mathbb{R})$ satisfying

$$\cdots, \subset V_{-2}, \subset V_{-1}, \subset V_0, \subset V_1, \subset V_2, \subset, \cdots,$$

with "difference space" $W_j = V_{j+1} \ominus V_j$ where $j = \cdots, -1, 0, 1, \cdots$. When a sequence of subspaces satisfies MRA properties (see Vidakovic 1999), there exists an orthonormal basis $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ with integers j and k for $L^2(\mathbb{R})$. In other words, the orthonormal wavelet basis ψ_{jk} is obtained by dilating and translating a *mother wavelet* ψ. For $L^2(\mathbb{R}) = \bigoplus_{j=-\infty}^{\infty} W_j$, a function f can then be represented by the wavelet series

$$f(x) = \sum_{j,k} d_{jk}\psi_{jk}(x), \tag{2.1}$$

with wavelet coefficients $d_{jk} = \langle f, \psi_{jk} \rangle = \int f(x)\psi_{jk}dx$ describing features of the function f at the spatial locations indexed by k and scales indexed by j. Other "children" wavelets are generated by translations of the scaling function ϕ and dilations and translations of the mother wavelet ψ using the following relationships:

$$\begin{aligned}\phi_{j_0k}(t) &= 2^{j_0/2}\ \phi(2^{j_0}t-k),\\ \psi_{jk}(t) &= 2^{j/2}\ \psi(2^{j}t-k), \qquad j=j_0, j_0+1, \cdots;\ k\in\mathbb{Z}\end{aligned} \tag{2.2}$$

for some fixed $j_0 \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers.

Given the wavelet bases defined above, a function $f \in L^2(\mathbb{R}) = V_{j_0} \oplus \{\bigoplus_{j=j_0}^{\infty} W_j\}$ is represented in a corresponding wavelet series as:

$$f(t) = \sum_{k\in\mathbb{Z}} c_{j_0k}\phi_{j_0k}(t)\ +\ \sum_{j=j_0}^{\infty}\sum_{k\in\mathbb{Z}} d_{jk}\psi_{jk}(t), \tag{2.3}$$

where $c_{j_0k} = \langle f, \phi_{j_0k}\rangle$ and $d_{jk} = \langle f, \psi_{jk}\rangle$. Note that $\langle\cdot,\cdot\rangle$ is the standard L^2-inner product of two functions: $\langle g_1, g_2\rangle = \int_{\mathbb{R}} g_1(t)g_2(t)dt$.

2. Wavelet transformation and shrinkage

Wavelets have been extremely successful as a tool for the analysis and synthesis of discrete data. Let $\boldsymbol{y} = (y_1, \cdots, y_p)'$ be observations of a function taken at p equally spaced time points. I assume p to be a power of two for computational convenience in this chapter. A fast algorithm, the *discrete wavelet transformation* (DWT), exists for decomposing $\boldsymbol{y}$ into a set of p wavelet coefficients, Mallat (1989), in only $O(p)$ operations. A wavelet transformation with $j = 1, \cdots, J$ levels (or scales) can be seen as a cumulative measure of the variation in the data over regions proportional to the J scales, with coefficients at coarser levels, i.e. for increasing values of j, describing features at lower frequency ranges and larger time periods, Percival and Walden (2000). Although it operates in practice by means of linear recursive filters, the DWT can be also represented in matrix form as $\boldsymbol{d} = \boldsymbol{W}\boldsymbol{y}$ with $\boldsymbol{W}$ an orthogonal matrix corresponding to the discrete wavelet transform and $\boldsymbol{d}$ a vector of wavelet coefficients describing features of the function at the J scales. An algorithm for the inverse reconstruction, the *inverse DWT* (IDWT), also exists. There are many different wavelet families: Harr's, Shannon's, Meyer's, Franklin's and Daubechies'. In my research work, I choose to use Daubechies's wavelets. Daubechies (1992) first proposed a class of wavelet families which have compact support ensuring a good localization in time and maximum number of vanishing moments for any given smoothness. These properties allow an effective and parsimonious representation of functions with local behavior, which is why Daubechies wavelets are extensively used in applications. One will use the symbol *Daub#N*, $N \in \mathbb{N}$ for Daubechies wavelet with vanishing moment N. The mother wavelet of the family *Daub#N* has a compact support of length $2N - 1$ and N vanishing moments, that is ψ is orthogonal to all polynomials with order $\leq N$ such that

$$\int_{\mathbb{R}} x^k \psi(x) dx = 0, \ k = 0, 1, \cdots, N.$$

For larger N, one has wavelets with wider but more regular support.

Wavelet shrinkage refers to the estimation of a function from noisy observations and it is a well-known application of wavelets. Wavelet shrinkage usually refers to reconstructions obtained from the shrunk wavelet coefficients, which means that a wavelet transformation is applied to the data and the noise is removed by thresholding or shrinking the smallest wavelet coefficients. Donoho and Johnstone (1994, 1998) suggest a simple recipe for wavelet estimation by the following three steps.

Step 1. Transform the observations y_i, $i = 1, \cdots, n$ to the wavelet domain by applying a discrete wavelet transformation. The result is a sequence of wavelet coefficients d_i, $i = 1, \cdots, n$.

Step 2. Estimate σ, which is a standard deviation at noise level. Use this estimator to threshold (or shrink) the wavelet coefficients.

Step 3. Invert the thresholded (shrunk) coefficients, recovering the estimator of the function $\hat{f}_i$.

The most common thresholding rules are *hard* and *soft*, which are expressed as follows respectively:

$$\delta^h(d,\lambda) = \begin{cases} 0 & \text{if } \ |d| \le \lambda \\ d & \text{if } \ |d| > \lambda, \end{cases}$$

and

$$\delta^s(d,\lambda) = \begin{cases} 0 & \text{if } \ |d| \le \lambda \\ d-\lambda & \text{if } \ d > \lambda \\ d+\lambda & \text{if } \ d < -\lambda \end{cases}$$

where d indicates wavelet coefficients and $\lambda \ge 0, \ d \in \mathbb{R}$. As can be seen above, *hard threshold* rule "keep (keep the coefficients)" or "kill (replace 0 with coefficients)" wavelet coefficients whereas *soft thresholding* is a "shrink" or "kill" rule. The threshold λ should be estimated from data. Thus, thresholding allows the data itself to decide which wavelet coefficients are significant.

There are a variety of approaches to estimate the threshold level λ. They can be categorized into two groups: *global threshold* and *level-dependent threshold.* With the first group of thresholds, a single value λ would be applied to all wavelet coefficients. There also exist a level-dependent value λ_j for each resolution level $j = j_0, \cdots, J-1$. I will introduce two of them from each category: *universal threshold* and *SureShrink.* The threshold $\lambda = \sqrt{2\log n}\sigma$ is called *universal threshold* by Donoho and Johnstone (1994). There are several methods to estimate σ which is usually unknown. Most of the choices to estimate σ involve the wavelet coefficients at fine scales. Often, only the finest scale is used to estimate the variance of noise. The signal-to-noise ratio is usually small at high resolutions and, if the signal is not too irregular, the finest scale should contain mainly noise. Moreover, the finest scale contains 50% of all coefficients. Some standard estimators of σ are

$$\hat{\sigma} = \sqrt{\frac{1}{n/2-1}\sum_{i=1}^{n/2}[d_i^{(J-1)} - \bar{d}^{(J-1)}]^2},$$

where $\bar{d}^{(J-1)}$ is a mean of wavelet coefficients at level $J-1$, or a more robust MAD (median absolute deviation from the median),

$$\hat{\sigma} = 1.4826 \times \text{MEDIAN}[|\underset{\sim}{d}^{(J-1)} - \text{MEDIAN}(\underset{\sim}{d}^{(J-1)}|)]$$

where $\underset{\sim}{d}^{(J-1)}$ is the vector of finest detail coefficients.

The adaptation in *SureShrink* is achieved by specifying thresholds level-wise.

$$\text{SURE}(\boldsymbol{d}, \lambda) = k - 2\sum_{i=1}^{k} \boldsymbol{I}(|d_i| \le \lambda) + \sum_{i=1}^{k} \min(|d_i|, \lambda)^2 \tag{2.4}$$

The threshold level λ is set so as to minimize the estimate $\text{SURE}(\boldsymbol{d}, \lambda)$ for a given wavelet coefficients,

$$\lambda^{sure} = \arg \min_{0 \le \lambda \le \lambda^U} \text{SURE}(\boldsymbol{d}, \lambda). \tag{2.5}$$

The threshold λ^{sure} and the soft-thresholding rule are the core of the level dependent procedure Donoho and Johnstone call this *SureShrink*. If the wavelet representation at a particular level is not too sparse, the SURE threshold is used. Otherwise, the universal threshold is selected. The level j is considered sparse if

$$s_j^2 \le \frac{1}{\sqrt{n_j}} \log_2 n_j^{3/2},$$

where n_j is the number of coefficients in the j-th level, and $s_j^2 = \frac{1}{n_j}\sum_{k=1}^{n_j}(d_{jk}^2 - 1)$. It is common for all thresholding rules to set to 0 the coordinates of a vector $\boldsymbol{d}$, which is subjected to thresholding, if they are smaller in absolute value than the threshold λ.

Bayesian approaches have also been proposed that use mixture priors on the wavelet coefficients. The recent review paper of Antoniadis, Bigot and Sapatinas

(2001) provides an exhaustive review of the different approaches, classical and Bayesian, and related extensions. All these approaches are limited to the single function setting. Wavelet-based methods for the analysis of multiple curves are described by Brown, Fearn and Vannucci (2001) who considered regression models to relate a multivariate response to functional predictors, applied wavelet transforms to the curves, and used Bayesian selection methods to identify features that best predict the responses. Vannucci, Sha and Brown (2005) considered classification problems and extended wavelet methods to probit models. Also, Morris, Vannucci, Brown and Carroll (2003) applied ideas of wavelet regression to the setting of nested functional modelling.

VITA

Sinae Kim was born in 1976, in Daegu, Korea. She graduated from Dongnae girl's high school in Busan, Korea in 1995. She received a Bachelor of Science degree in statistics from Pusan National University in Busan, Korea in February 1999. She received a Master of Science in statistics from Texas A&M University in College Station, Texas under the direction of Professor Suojin Wang in May 2003. She continued her study in statistics under the direction of Professor Marina Vannucci, and received a Doctor of Philosophy degree in statistics from Texas A&M University in May 2006. Her permanent address is 90-8 Jangjeon1-Dong Keumjeong-Gu, Busan, Korea.