University of Tennessee, Knoxville

Trace: Tennessee Research and Creative Exchange

Masters Theses

Graduate School

5-2005

Novel Algorithms and Datamining for Clustering Massive Datasets

Aruna K. Buddana

University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Aruna K. Buddana entitled "Novel Algorithms and Datamining for Clustering Massive Datasets." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Statistics.

Halima Bensmail, Major Professor

We have read this thesis and recommend its acceptance:

Mary Leitnaker, Robert Mee

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Aruna K. Buddana entitled ”Novel Algorithms and Datamining for Clustering Massive Datasets”. I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Statistics.

Halima Bensmail
Major Professor

We have read this thesis and
recommend its acceptance:

Mary Leitnaker

Robert Mee

Accepted for the Council:

Anne Mayhew
Vice Chancellor and
Dean of Graduate Studies

(Original signatures are on file with official student records.)

Novel Algorithms and Datamining for Clustering Massive Datasets

A Thesis
Presented for the
Master of Science Degree
The University of Tennessee, Knoxville

Aruna K Buddana
May 2005

Dedicated to my family, teachers and friends

ACKNOWLEDGEMENTS

Foremost, I would like to express my deepest gratitude to my advisor, Dr. Halima Bensmail, for her excellent guidance and enormous support during my graduate study at The University of Tennessee, Knoxville. No words of gratitude are sufficient to mention here, to describe her unrelenting support.

Many thanks to Dr. Mary Leitnaker and Dr. Robert Mee for serving on my thesis committee. I am grateful to Dr. George Ostrouchov, Senior Research Staff Member, ORNL , for giving me the opportunity to work as an intern and gain research experience.

I am deeply indebted to my parents for the support and the motivation they provided me to explore higher levels of education. I thank all my buddies especially Jennifer Golek for their endless support. Nothing would be possible without their good wishes. Last, and surely not the least, my heart full thanks to my dearest friend, my fiancée, Suresh, for his motivation and being there for me.

ABSTRACT

Clustering proteomics data is a challenging problem for any traditional clustering algorithm. Usually, the number of samples is much smaller than the number of protein peaks. The use of a clustering algorithm which does not take into consideration the number of feature of variables (here the number of peaks) is needed. An innovative hierarchical clustering algorithm may be a good approach. This work proposes a new dissimilarity measure for the hierarchical clustering combined with a functional data analysis. This work presents a specific application of functional data analysis (FDA) to a highthrouput proteomics study. The high performance of the proposed algorithm is compared to two popular dissimilarity measures in the clustering of normal and Human T Cell Leukemia Virus Type 1 (HTLV-1)-infected patients samples.

The difficulty in clustering spatial data is that the data is multi - dimensional and massive. Sometimes, an automated clustering algorithm may not be sufficient to cluster this type of data. An iterative clustering algorithm along with the capability of visual steering may be a good approach. This case study proposes a new iterative algorithm which is the combination of automated clustering methods like the bayesian clustering, detection of multivariate outliers, and the visual clustering. Simulated data from a plasma experiment and real astronomical data are used to test the performance of the algorithm.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The entire work can be broadly divided into two categories:

1) Hierarchical Clustering High Dimensional Proteomics Data using Functional Data Analysis

2) Clustering Massive Spatial Data using Iterative Clustering Algorithm

1.1 Proteomics Data

A variety of mass spectrometry-based platforms are currently available for providing information on both protein patterns and protein identity [1, 2]. Specifically, the first widely used such mass spectrometric technique is known as surface enhanced laser desorption ionization (SELDI) coupled with time of flight (TOF) mass spectrometric detection [3 – 5]. The SELDI approach is based on the use of an energy-absorbing matrix such as sinapinic acid (SPH), large molecules such as peptides ionize instead of decomposing when subjected to a nitrogen UV laser. Thus, partially purified serum is crystallized with an SPH matrix and placed on a metal slide. Depending upon the range of masses the investigator wishes to study, there are a variety of possible slide surfaces; for example, the strong anion exchange (SAX) or the weak cation exchange (WCX) surface. The peptides are ionized by the pulsed laser beam and then traverse a magnetic field-containing column. Masses are separated according to their times of flight as the latter are proportional to the square of the mass-to-charge (m/z) ratio. Since nearly all of the resulting ions have unit charge, the mass-to-charge ratio is in most cases a mass. The spectrum (intensity level as a function of mass) is recorded, so the resulting data obtained on each serum sample are a series of intensity levels at each mass value on a common grid of masses (peaks).

Proteomic profiling is a new approach to clinical diagnosis, and many computational challenges still exist. Not only are the platforms themselves still improving, but the methods used to interpret the high dimensional data are developing as well [6, 7].

A variety of clustering approaches has been applied to high dimentional genomics and proteomics data [8 – 11]. Hierarchical clustering methods give rise to nested partitions, meaning the intersection of a set in the partition at one level of the hierarchy with a set of the partition at a higher level of the hierarchy will always be equal to the set from the lower level or the empty set. The hierarchy can thus be graphically represented by a tree.

Functional Data Analysis (FDA) is a statistical data analysis represented by smooth curves or continuous functions $\mu_i(t)$; $i = 1,..,n,$ where n is the number of observations and t might or might not necessarily denote time but might have a general meaning. Here t denotes the mass (M/Z). In practice, the information over $\mu_i(t)$ is collected at a finite number of points, T_i, thus observing the data vector $\mathbf{y}_i = (y_{i1,}\ldots,y_{iT_i})^t$. The basic statistical model of FDA is given by

$$y_{ij} = \hat{\mu}_i(t_{ij}) = \mu_i(t_{ij}) + \epsilon_i(t_{ij}) \text{ or } \quad i = 1,\ldots,n; j = 1,..,T_i \tag{1}$$

where t_{ij} is the mass value at which the j^{th} measurement is taken for the i^{th} function $\boldsymbol{\mu}_i$. The independent disturbance terms $\epsilon_i(t_{ij})$ are responsible for roughness in $\mathbf{y}_i$.

FDA has been developed for analyzing functional (or curve) data. In FDA, data consists of functions not of vectors. Samples are taken at time points $t_1, t_2,\ldots,$ and regard $\mu_i(t_{ij})$ as multivariate observations. In this sense the original functional y_{ij} can be regarded as the limit of $\mu_i(t_{ij})$ as the sampling interval tends to zero and the dimension of multivariate observations tends to infinity. Ramsay and Silverman [12, 13] have discussed several methods for analyzing functional data, including functional regression analysis, functional principal component analysis (PCA), and functional canonical correlation analysis (CCA). These methodologies look attractive, because one often meets the cases where one wishes to apply regression analysis and principal component analysis to such data. In the following we describe how to use the FDA tools for applying functional data analysis and a new disssimilarity measure to classify the spectra data.

We propose to implement a hierarchical clustering algorithm for proteomics data using functional data analysis. We use functional transformation to smooth and reduce the dimensionality of the spectra and develop a new algorithm for clustering high dimensional proteomics data.

1.2 Spatial Data

Spatial Databases are the database systems for the organization of spatial data i.e. the point objects extended in a 2-Dimensional or a 3-Dimensional or some other higher dimensional vector space. Multi-scale databases are a set of spatial databases with certain limitations

Data mining, an essential element of much detailed process Knowledge Discovery in Databases (KDD), deals with the extraction of hidden structure of huge data sets either by clustering or by discovering regularities in the data. The subject of data mining spatial databases finds its roots from the concept of data mining relational and transactional databases. Knowledge discovery in large multi-scale databases has become essential since data collected from many applications such as satellite images, astrophysical equipment, X-ray crystallography is stored in such massive spatial databases. Also, spatial data mining techniques have wide applications ranging from remote sensing, geographical information systems to crime scene investigations and environment & planning. Because of the large volumes of the data in these multi-scale databases it is often expensive and unrealistic to look at every detail for any information buried in the dataset.

Clustering analysis is a renowned data mining technique, which involves dividing a large dataset into meaningful subclasses and thus extracting hidden patterns among the objects. It is a procedure to extract the gist of information present in the data set. It usually demands some information to be known such as the statistical distribution of the data (if the data is gaussian distributed) and number of clusters one can expect. But real time spatial data may not have any of this information available. Also the shape of the clusters can be very arbitrary such as spherical, linear, ellipsoidal, elongated etc.

These clusters can be populated with as many as 100,000 points or as few as 10 points in a given time. So uniform generalized cluster analysis is almost unimaginable in massive multi-scale data mining.

Before we actually delve into spatial data mining, a brief background of the clustering methods is presented in the next section.

1.3 Clustering Methods

The term cluster analysis (first used by Tryon, 1939) includes a number of different algorithms and routines for grouping similar objects into particular categories.

The clustering algorithms are defined to be the procedures, which produce clusters of data from a given dataset.

Clustering algorithms can be broadly classified into hierarchical clustering techniques and optimization partitioning algorithms. The hierarchical algorithms operate in such a way that the dataset is divided into certain groups sequentially making all the objects similar for a given branch node until higher up the tree. These algorithms can be further split into agglomerative and splitting procedures. In agglomerative technique, hierarchical clustering starts from the optimum partition possible (each observation forms a cluster) and groups them. This procedure depends on the definition of the distance between two clusters. Single linkage, complete linkage, average linkage and Ward distance are frequently used distances. Splitting procedure starts with the crudest partition possible one cluster contains all of the observations. It proceeds by splitting the single cluster up into smaller sized clusters.

Divisive methods are not generally available, and rarely have been applied.

The partioning algorithms divide the dataset into homogenous clusters until a certain score is optimized. The most common algorithm is the k-means algorithm in this category. Like most other clustering algorithms k-means does not necessarily find an optimized configuration. It performs differently on different datasets and it is more biased on the selection of the initial random clusters.

The main difference between the two clustering techniques is that in hierarchical clustering once groups are found and elements are assigned to the groups, this assignment cannot be changed. In partitioning techniques, on the other hand, the assignment of objects into groups may change during the algorithm application.

Clustering algorithms face their toughest problem in implementation when the dataset is massive and multi dimensional. The hierarchical clustering methods cannot compute their distance matrices on the basis of which clustering is done. The k-means algorithm becomes limited once the dataset is considerably large.

CHAPTER 2
MATERIALS

2.1 Serum Samples

Protein expression profiles generated through SELDI analysis of sera from HTLV-1 (Human T cell Leukemia virus type 1)-infected individuals were used to determine the changes in the cell proteome that characterize Adult T cell leukemia (ATL), an aggressive lymphoproliferative disease from HTLV-1-Associated Myelopathy/Tropical Spastic Paraparesis (HAM/TSP), a chronic progressive neurodegenerative disease. Both diseases are associated with the infection of T-cells by HTLV-1. The HTLV-1 virally encoded oncoprotein Tax has been implicated in the retrovirus mediated cellular transformation and is believed to contribute to the oncogenic process through induction of genomic instability affecting both DNA repair integrity and cell cycle progression [14, 15]. Serum samples were obtained from the Virginia Prostate Center Tissue and body fluid bank. All samples had been procured from concenting patients according to protocols approved by the Institutional Review Board and stored frozen. None of the samples had been thawed more than twice.

Triplicate serum samples ($n = 68$) from healthy or normal ($n_1 = 37$), ATL ($n_2 = 20$) and HAM ($n_3 = 11$) patients were processed. A bioprocessor, which holds 12 chips in place, was used to process 96 samples at one time. Each chip contained one ”QC spot” from normal pooled serum, which was applied to each chip along with the test samples in a random fashion. The QC spots served as quality control for assay and chip variability. The samples were blinded for the technicians who processed the samples. The reproducibility of the SELDI spectra, i.e., mass and intensity from array to array on a single chip (intraassay) and between chips (interassay), was determined with the pooled normal serum QC sample [Figure 1].

2.2 SELDI Mass Spectrometry

Serum samples were analyzed by SELDI mass spectrometry as described earlier [16]. The spectral data generated was used in this study for the development of the novel functional data analysis.

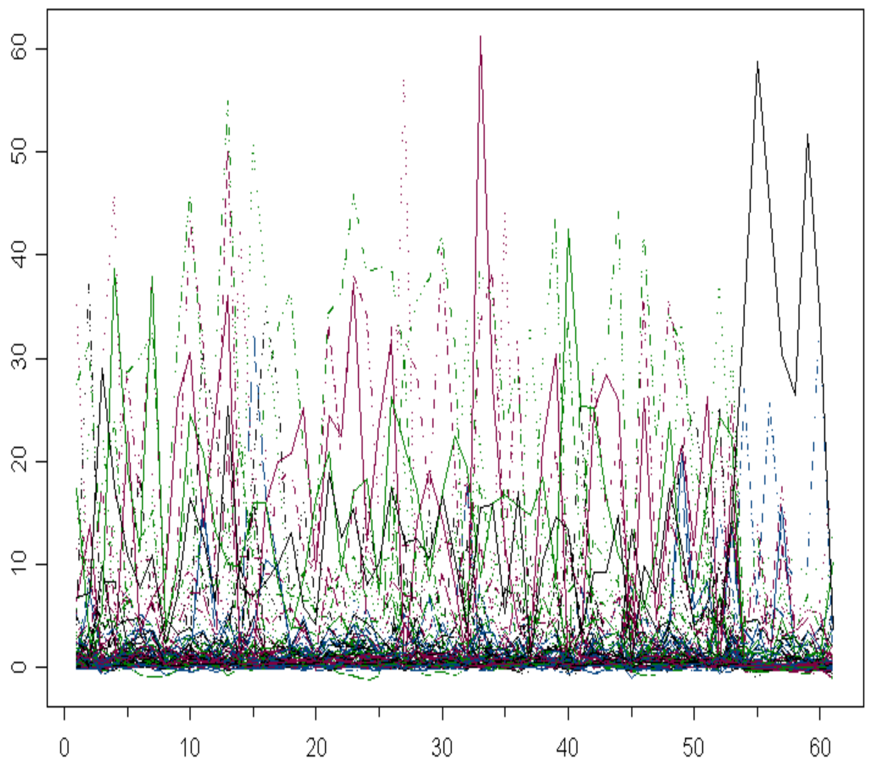


Figure 1: Three cut-expressions from a normal, HAM and an ATL patient

2.3 Plasma Experiment

Because of an ever-expanding global population, the world's demand for energy grows more and more. Energy sources that come from burning fossil fuels contribute to environmental stress and release of greenhouse gasses. Nuclear fusion is an alternative energy source with the potential to provide abundant clean energy. Unlike current nuclear fission power plants that are based on splitting atoms heavy elements, nuclear fusion is a process that combines light elements. Unlike fission processes, nuclear fusion produces no harmful waste. Tokamak facilities [Figure 2] are experimental test chambers used for testing nuclear fusion energy, where Hydrogen nuclei can be fused into Helium, mimicking the process that occurs at the center of the Sun. The common type of tokamak fusion labs have doughnut-shaped chambers where the fusion reaction occurs. In order to achieve nuclear fusion, the plasma inside the chamber must be heated to enormous temperatures. Plasma is the fourth state of matter (solid, liquid, gas, plasma). It is also the most common state of matter, making up 99% of the visible universe. Best known examples of plasma include flames, lightning, neon signs and fluorescent lights, and the aurora borealis. Tokamak chambers apply radio-frequency heating to drive the plasma to high temperatures in similar way that microwave ovens heat water. They also employ powerful magnetic fields to contain the plasma inside the vessel. The largest current tokamak is the Europe's JET facility (www.jet.efda.org). Larger tokamaks are planned, including ITER (www.iter.org), a multinational venture that will heat plasma to 100 million ^{0}C and produce 500 Megawatts of power. ITER is projected to be the progenitor of a commercially viable fusion energy source.

Because the construction of a tokamak facility is an expensive enterprise, and the processes governing the behavior of the tokamak are highly complex and multivariate, researchers rely on computer models and simulations to guide their designs. For our project, the data was obtained from the All-Orders Spectral Algorithm (AORSA) computer model for electromagnet wave interaction with magnetized plasmas that enables physics insights and a quantitative understanding of a wide range of radio frequency - plasma interactions Jaeger et al. (2002).

The spatial data is obtained from the All-Orders Spectral Algorithm (AORSA) computer model for electromagnet wave interaction with magnetized plasmas that enables physics insights and a quantitative understanding of a wide range of radio frequency - plasma interactions Jaeger et al. (2002). Plasma is the fourth state of matter (solid, liquid, gas, plasma). It is also the most common state of matter, making up 99% of the visible universe. Best known examples of plasma include flames, lightning, neon signs and fluorescent lights, and the aurora borealis.

A typical AORSA computational experiment at the Oak Ridge National Laboratory (ORNL) runs for many hours on a large parallel supercomputer and produces large amounts of data.

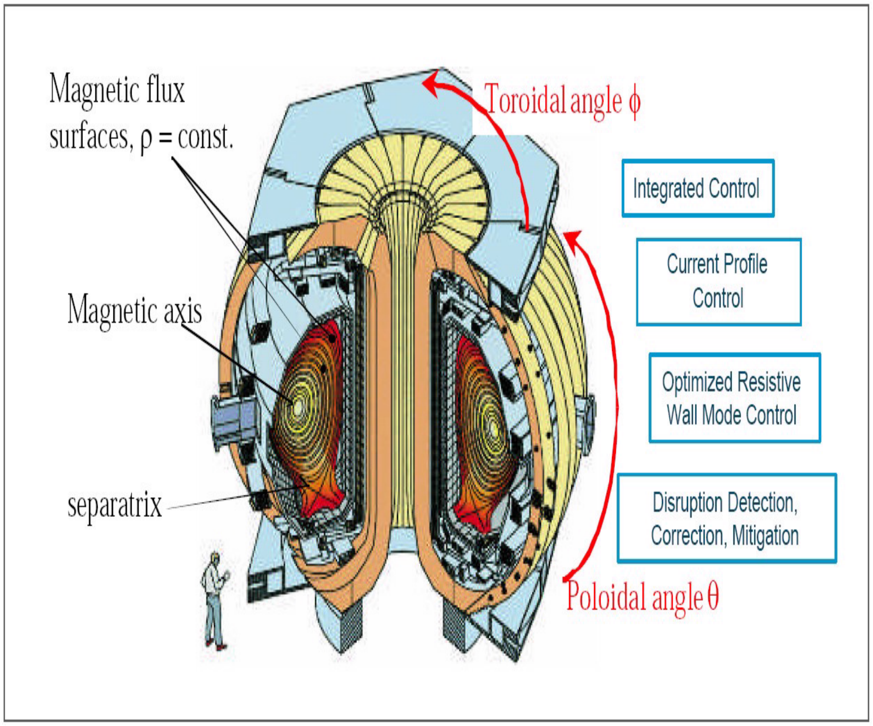

Magnetic flux surfaces, ρ = const.
Toroidal angle ϕ
Integrated Control
Current Profile Control
Magnetic axis
Optimized Resistive Wall Mode Control
separatrix
Disruption Detection, Correction, Mitigation
Poloidal angle θ

Figure 2: Integrated simulation & optimization of fusion systems

We selected a subset of the data that describes radio frequency heating of tokamak plasma. A tokamak is a toroidal vessel (imagine a fat doughnut) where electromagnetically confined plasma is radio frequency heated to fusion temperatures. Our data set consists of four quasi-linear diffusion coefficients, b, c, e, and f, with units of $\frac{v^2}{s}$(velocity squared per second). Coefficient values are obtained by averaging around tubes of radius ρ for given $u_\perp$ (perpendicular velocity) and $u_\|$(parallel velocity). Imagine doughnut-shaped shells of tube radius ρ and binned according to perpendicular and parallel plasma velocities ($u_\perp \times u_\|$). We cluster the four diffusion coefficients by pooling data over $\rho \times u_\perp \times u_\|$ $(32 \times 65 \times 129)$ for a total of $268,320$ observations. The spatial dimensions $\rho \times u_\perp \times u_\|$ can be used to map the resulting clusters and interpret the mapped results.This data set was chosen because it represents a particularly difficult situation for an automated clustering algorithm while providing visually stunning clusters. It is a difficult data set not only because it is large but also because cluster sizes range across several scales. Here we consider clustering the diffusion coefficient data and leave the mapping in the $\rho \times u_\perp \times u_\|$ space to a separate project.

CHAPTER 3
METHODS

3.1 Hierarchical Clustering Using FDA

We propose to implement a hierarchical clustering algorithm for proteomics data using functional data analysis, which consists of detecting hidden group structures within a functional data set. We apply a new dissimilarity measure to the smoothed (transformed) proteomics functions $\hat{\mu}_i$. Then we develop a new metric that calculates the dissimilarity between different curves produced by protein expression. The development of metrics for curve and time series models was first addressed by Piccolo and Corduas (1990). Heckman and Zamar proposed a dissimilarity measure δ_{HZ} for clustering curves (2000). Their dissimilarity measure considers curve invariance under monotone transformations. Let $\Lambda_i = \left\{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{m_i}^{(i)}\right\}$ be the collection of the estimated points where the curve $\boldsymbol{\mu}_i(t)$ has a local maximum and let m_i be the number of maximal per observation or per sample (i). δ_{HZ} is defined as:

$$\delta_{HZ}(i,l) = \frac{\sum_{j=1}^{m_i}\left(r(\lambda_j^{(i)}) - \overline{r(\lambda^{(i)})}\right)\left(r(\lambda_j^{(l)}) - \overline{r(\lambda^{(l)})}\right)}{\sum_{j=1}^{m_i}\left(r(\lambda_j^{(i)}) - \overline{r(\lambda^{(i)})}\right)^2 \sum_{j=1}^{m_l}\left(r(\lambda^{(l)}) - \overline{r(\lambda^{(l)})}\right)^2}$$

where

$$\begin{aligned} r(\lambda_j^{(i)}) &= k_j^{(i)} + u_j^{(i)}/2, \\ k_j^{(i)} &= \{\#\, i,\ \lambda_i^{(i)} < \lambda_j^{(i)}\}, \\ u_j^{(i)} &= \{\#\, i,\ \lambda_i^{(i)} = \lambda_j^{(i)}\} \\ \overline{r(\lambda^{(i)})} &= \frac{1}{m_i}\sum_{j=1}^{m_i} r(\lambda_j^{(i)}) \end{aligned}$$

This measure is powerful for regression curves which are mainly monotone. On the other hand, Cerioli et al. (2003), propose a dissimilarity measure δ_C extending the one proposed by Ingrassia et al. (2003). Cerioli's dissimilarity δ_C is defined by:

$$d(i,l) = \sum_{j=1}^{m_i} \frac{|\lambda_j^{(i)} - \lambda_{*j}^{(l)}|}{m_i}, \; \lambda_{*j}^{(l)} = \left\{\lambda_{j'}^{(l)} : |\lambda_j^{(i)} - \lambda_{j'}^{(l)}| = \min,\; i = 1,..,n\right\}$$

$$\delta_C(i,l) = \left(\frac{d_{il} + d_{li}}{2}\right)$$

Both dissimilarity measures shows good performance for time series data. Dissimilarity δ_C does not involve all the indices m_i of the smoothed curve. It also uses the shortest distance between curves by involving few data points obtained by FDA smoothing.

A flexible dissimilarity measure is the one that may combine the characteristic of both measures δ_{HZ} and δ_C. This means that a potential dissimilarity measure should use the collected estimated points of the original curve obtained from FDA so that no information is lost and should work on different type of smoothed curves without using the monotonicity restriction.

In this sense, we propose a functional-based dissimilarity δ_B measure which uses the rank of the curve proposed by Heckman and Zamar and generalizes Cerioli et al. dissimilarity measure as the following:

$$d_{il} = \sum_{j=1}^{m_i} \frac{|r(\lambda_j^{(i)}) - r(\lambda_{*j}^{(l)})|}{m_i},$$

$$r(\lambda_{*j}^{(l)}) = \frac{\sum_{h=1}^{m_l} |r(\lambda_j^{(i)}) - r(\lambda_{h'}^{(l)})|}{m_l}$$

$$r(\lambda_j^{(i)}) = k_j^{(i)} + u_j^{(i)}/2 \text{ and } k_j^{(i)} = \{\# \, i,\; \lambda_i^{(i)} < \lambda_j^{(i)}\}$$

$$u_j^{(i)} = \{\# \, i,\; \lambda_i^{(i)} = \lambda_j^{(i)}\} \text{ and } \overline{r(\lambda^{(i)})} = \frac{1}{m_i} \sum_{j=1}^{m_i} r(\lambda_j^{(i)})$$

Obviously, $d_{ii} = 0$ and $d_{il} = 0$, if μ_i and μ_l have the same shape ($T_i = T_l$). We can adjust the formula above to obtain a dissimilarity measure that satisfies symmetry, by taking δ_B as our proposed dissimilarity measure:

$$\delta_B(i,l) = \left(\frac{d_{il} + d_{li}}{2}\right)$$

We used three powerful hierarchical methods to derive clusters or patterns using δ_B and we compare the performance of δ_B to δ_C and δ_{HZ}. The hierarchical algorithms we used are (1) **Pam** which partitions the data into different clusters "around their medoids", (2) **Clara** works as in 'Pam'. Once the number of clusters is specified and representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid. The sum of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the sum is minimal, is retained. Each sub-dataset is forced to contain the medoids obtained from the best sub-dataset until then, (3) **Diana** is probably unique in computing a divisive hierarchy, whereas most other software for hierarchical clustering is agglomerative. Moreover, 'Diana' provides the divisive coefficient which measures the amount of clustering structure found. The 'Diana'-algorithm constructs a hierarchy of clustering starting with one large cluster containing all n observations. Clusters are divided until each cluster contains only a single observation. At each stage, the cluster with the largest diameter is selected.

3.2 Model-Based Cluster Analysis

In cluster analysis, we consider the problem of determining the structure of the data with respect to clusters when no information other than the observed values is available; from the extensive literature, we mention Hartigan (1975), Gordon (1999), and Kaufman and Rousseeuw (1990). Important references on the statistical aspects of cluster analysis include MacQueen (1967) , Wolfe(1978), Scott and Symons (1971), and Bock(1985). Various strategies for simultaneous determinating of the number of clusters and the cluster membership have been proposed (e.g. Engelman and Hartigan (1969)); Bozdogan 1993), for a review see Bock (1996). An alternative is described in this paper based on the reparameterization of the covariance matrices using a fully Bayesian framework.

Mixture models provide a useful statistical frame of reference for cluster analysis. The Bayesian approach is promising for a variety of mixture models, both Gaussian and non Gaussian (Binder, 1981; Banfield and Raftery, 1993; McLachlan and Peel, 2000, Ch. 4).

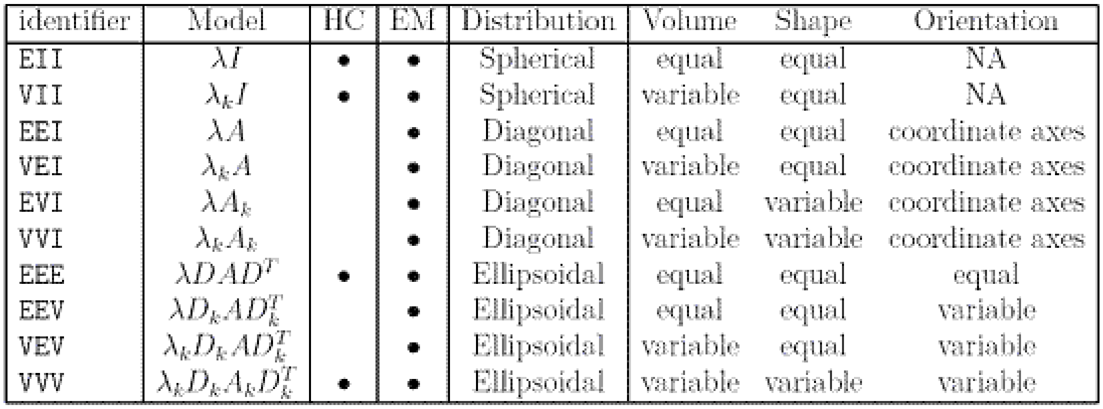
Banfield and Raftery (1993) –hereafter BR– introduced a new approach to cluster analysis based on a mixture of multivariate normal distributions, where the covariance matrices $\mathbf{\Sigma}_k$ in the classes are modelled in a geometrically interpretable way. Their approach is based on a variant of the standard spectral decomposition of $\mathbf{\Sigma}_k$, namely

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t \qquad (3)$$

where λ_k is a scalar, $\mathbf{A}_k = diag(1, a_{k2}, \ldots, a_{kp})$ where $1 \geq a_{k2} \geq \ldots a_{kp} > 0$, and $\mathbf{D}_k$ is an orthogonal matrix for each $k = 1, \ldots, K$.

Bensmail, Celeux, Raftery and Robert (1997) proposed a Bayesian approach which overcomes the limitations mentioned above ((a),....,(g)). However, only four models for $\mathbf{\Sigma}_k$ were explicitely considered. These are the spherical models $[\lambda]$ and $[\lambda_k \mathbf{I}]$ (in what follows, [.] is used to indicate a particular model for $\mathbf{\Sigma}_k$), the linear model $[\mathbf{\Sigma}]$ and the proportional model $[\lambda_k \mathbf{\Sigma}]$. Dasgupta and Raftery (1998) used the model $[\mathbf{D}_k \mathbf{A} \mathbf{D}_k^t]$ to detect features in a spatial point process where the shape matrix **A**

Table 1: Different geometric models



identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
EII	λI	●	●	Spherical	equal	equal	NA
VII	$\lambda_k I$	●	●	Spherical	variable	equal	NA
EEI	λA		●	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k A$		●	Diagonal	variable	equal	coordinate axes
EVI	λA_k		●	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k A_k$		●	Diagonal	variable	variable	coordinate axes
EEE	λDAD^T	●	●	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k A D_k^T$		●	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k D_k A D_k^T$		●	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k^T$	●	●	Ellipsoidal	variable	variable	variable

was unknown but they constrained the diagonal terms of the shape matrix to be equal: $\mathbf{A} = diag\{1, \alpha, \ldots, \alpha\}$ and to have a low value.

Our particular interest is to extend this previous work in two respects, using the fully Bayesian inference we develop here: first, to the family of models where the covariance matrix $\mathbf{\Sigma}_k$ is represented by $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$, $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^t$ and $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^t$ respectively, where all the parameters $\lambda, \lambda_k, \mathbf{A}, \mathbf{A}_k$ and $\mathbf{D}_k$ $(k = 1, \ldots, K)$ are unknown; second, to the case where the data contain outliers. The parameters involved in the parameterization of the covariance matrix $\mathbf{\Sigma}_k$ are unknown and not constrained as in Banfield and Raftery (1993), and Dasgupta and Raftery (1998). The models we are discussing here can be applied to minefields and seismic faults, earthquake and other particular problems as discussed by Dasgupta and Raftery (1998). Table 1 shows the geometric interpretation of the various parametrizations used. We estimate $\boldsymbol{\pi}$, $\boldsymbol{\nu}$ and the parameters of the models given in Table 1 using Gibbs sampling.

We use the Laplace-Metropolis approximation to calculate the Bayes factor (Bensmail et al. 1997 and Dasgupta and Raftery 1998); the latter is used to choose the model and determine the number of groups simultaneously.

3.2.1 Bayesian cluster analysis: bclust

We assume that the data are generated by a mixture of underlying probability distributions; each component of the mixture represents a different cluster so that the observations $\mathbf{x}_i$ $(i = 1, \ldots, n; \mathbf{x}_i \in R^p)$ to be classified arise from a random vector X with likelihood density $p(\boldsymbol{\theta}, \boldsymbol{\pi} | X = \mathbf{x})$ as in (2), where $f_k(.|(\boldsymbol{\theta}_k = \boldsymbol{\mu}_k, \mathbf{\Sigma}_k))$ is the multivariate normal density function, $\boldsymbol{\mu}_k$ is the mean and $\mathbf{\Sigma}_k$ is the covariance matrix for the kth group. $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$ is the mixing proportion ($\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$). We are concerned with Bayesian inference about the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ and the classification indicators $\boldsymbol{\nu}$. Markov Chain Monte Carlo (MCMC) methods (e.g. Gilks, Richardson and Spiegelhatler, 1996) provide an efficient and general recipe for Bayesian analysis of mixtures. In fact, as explained in Gelman, Carlin, Stern and Rubin (1995) the key to Markov chain simulation is to create a Markov process whose stationary distribution is a specified $p(\boldsymbol{\theta}|\mathbf{x})$ and run the simulation long enough that the distribution of the current draws is close enough to the stationary distribution.When, as in our case, the posterior conditional distribution of the parameters is a complicated function of the parameters which in most cases are of high dimension, the MCMC algorithm is used to simulate a sample from the posterior distribution of each parameter and after convergence, the posterior mode of each sample is used as the Bayes estimate of the parameter considered. For instance, many authors have used the Gibbs sampler or the Data Augmentation method of Tanner and Wong (1987) (Wei and Tanner 1990 and Green 1995) for estimating parameters in univariate and multivariate Gaussian mixtures. One important consideration regarding the implementation of both algorithms is monitoring convergence. Tierney (1994) proved that both algorithms converge in probability to the true posterior distribution of the mixture parameters. The models we are investigating in this paper are described in Table 1.

Given a classification vector $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_n)$, we use the notation $n_k = \#\{i : \nu_i = k\}$ for the number of observations in cluster k, $\bar{\mathbf{x}}_k = \sum_{i:\nu_i=k} \mathbf{x}_i/n_k$ for the sample mean vector of all observations in cluster k, and $\mathbf{W}_k = \sum_{i:\nu_i=k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t$ for the sample covariance matrix. We use conjugate priors for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ of the mixture model. The prior distribution of the mixing proportions is a Dirichlet distribution

$$(\pi_1,\ldots,\pi_K) \sim \mathit{Dirichlet}\,(\alpha_1,\ldots,\alpha_K),$$

$$\left(\text{with joint distribution } p(\boldsymbol{\pi}) = \frac{\Gamma(\alpha_1+\ldots+\alpha_K)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)}\pi_1^{\alpha_1-1}\ldots\pi_K^{\alpha_K-1}\right)$$

The prior distributions of the means $\boldsymbol{\mu}_k$ of the mixture components conditionally on the covariance matrices $\boldsymbol{\Sigma}_k$ are Gaussian

$$(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) \sim N_p(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k/\tau_k). \tag{4}$$

with known scale factors $\tau_1,\ldots,\tau_K > 0$ and locations $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K \in R^p$, and in addition

$$\boldsymbol{\pi}, \boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_{\mathbf{K}} \text{ are independent}$$

$$\boldsymbol{\Sigma}_{\mathbf{1}},\ldots,\boldsymbol{\Sigma}_{\mathbf{K}}|\boldsymbol{\pi}, \boldsymbol{\mu}_{\mathbf{1}},\ldots,\boldsymbol{\mu}_{\mathbf{K}} \text{ are independent under different models}$$

The conjugate prior distribution of the covariance matrices $\boldsymbol{\Sigma}_k$ depends on the model, and will be given for each model in turn.

We estimate the parameters of the models in [Table 1] by determining the configurations $\boldsymbol{\pi}$, $\boldsymbol{\theta}$ that maximize the posterior density of $\boldsymbol{\pi}, \boldsymbol{\theta}|\mathbf{x}$ (posterior mode values). This posterior density is calculated (approximated) by the Gibbs sampler by simulating from the joint posterior distribution of $\boldsymbol{\pi}$, $\boldsymbol{\theta}$ and $\mathbf{v}$. At iteration $(t+1)$, the Gibbs sampler steps go as follows:

1. Simulate the classification variables $v_i^{(t+1)}, i = 1,\ldots,n$, independently according to the posterior probabilities $p_{ik}(\boldsymbol{\pi},\boldsymbol{\theta}) = P(v_i^{(t)} = k|\boldsymbol{\pi},\boldsymbol{\theta},\mathbf{x}_i)$ $(\mathbf{k} = \mathbf{1},\ldots,\mathbf{K})$ conditional on the current values for $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$ such that
$$p_{ik}^{(t+1)} = \pi_k f_k(\mathbf{x}_i|\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}) / \sum_{k=1}^{K} \pi_k^{(t)} f_k(\mathbf{x}_i|\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}) \qquad i = 1,\ldots,n,\ k = 1,\ldots,K.$$
There might be classes k which are empty. To solve this problem, we assign the observation which is closest to $\boldsymbol{\mu}_k^{(t)}$ to this empty class.
2. Simulate the vector $\boldsymbol{\pi}^{(t+1)} = (\pi_1^{(t+1)},\ldots,\pi_K^{(t+1)})$ of mixing proportions from its posterior distribution given $\mathbf{v}^{(t+1)}$, in particular from
$$\boldsymbol{\pi}^{(t+1)} \sim \mathit{Dirichlet}\ (\alpha_1 + n_1^{(t+1)},\ldots,\alpha_K + n_K^{(t+1)})$$
with α_k the known parameters of the prior Dirichlet distribution.
3. Simulate the parameter $\theta^{(t+1)}$ of the model from the posterior distribution $\boldsymbol{\theta}|\mathbf{v}^{(t+1)},\boldsymbol{\pi}$.
4. Iterate the steps 1 to 3.
(Details on the simulation of the parameters $\lambda, \lambda_k, \mathbf{A}, \mathbf{A}_k$ and $\mathbf{D}_k$ are discussed in the paragraphs of the Appendix).

The validity of this procedure, namely the fact that the Markov chain associated with the algorithm converges in distribution to the true posterior distribution of $\boldsymbol{\theta}$, was demonstrated by Diebolt and Robert (1994) in the context of one-dimensional normal mixtures. Their proof is based on a *duality principle*, which uses the finite space nature of the chain associated with the v_i's. This chain is ergodic with state space $\{1,\ldots,K\}$, and is thus geometrically convergent. These properties transfer automatically to the sequence of simulated values of θ and π, and important properties as the central limit theorem or the law of the iterated logarithm are then satisfied (Diebolt and Robert 1994).

For the models 1, 2, 3 and 4 of Table 1 the calculations are given in Bensmail et al. (1997), so we proceed here with the models 5-7. In what follows, we will describe the simulation steps in step 3 of the algorithm for the parameters to be estimated which are $\boldsymbol{\mu}_k$, λ, $\mathbf{A}$ and $\mathbf{D}_k$ $(k = 1,\ldots,K)$ for the model $[\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^t]$, $\boldsymbol{\mu}_k$, λ_k, $\mathbf{A}$ and $\mathbf{D}_k (k = 1,\ldots,K)$ for the model $[\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^t]$, and $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ $(k = 1,\ldots,K)$ for the general model $[\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^t]$.

(a) Model $[\lambda D_k A D_k^t]$

If the prior distribution of the parameter μ_k is as given in (4) and if the prior distribution of λ is assumed to be an inverse gamma distribution

$$\lambda \sim IG(\frac{m_0}{2},\frac{s_0}{2}) \text{ with density } p(\lambda) = \frac{(s_0/2)^{\frac{m_0}{2}}}{\Gamma(\frac{m_0}{2})}\lambda^{-(\frac{m_0}{2}+1)}e^{-\frac{s_0}{2\lambda}}$$

with m_0 and s_0 hyperparameters chosen by the user (see Section 5), then the posterior distribution of $(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k,\boldsymbol{\nu})$ is a multivariate normal distribution with mean $\bar{\xi}_k = (n_k\bar{\mathbf{x}}_k + \tau_k\xi_k)/(n_k + \tau_k)$ and covariance matrix $\boldsymbol{\Sigma}_k/(n_k + \tau_k)$. The posterior distribution of $\lambda|\mathbf{A},\mathbf{D}_1,\ldots,\mathbf{D}_K,\boldsymbol{\nu}$ is then given by

$$IG\left(\frac{m_0+np}{2},\frac{1}{2}\left\{s_0+\sum_k \mathrm{tr}\{\mathbf{D}_k\mathbf{A}^{-1}\mathbf{D}_k^t((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+\mathbf{W}_k+\boldsymbol{\Psi}_0)\}\right\}\right).$$

For the other parameters, we assume that

$$\boldsymbol{\Sigma}_k \sim W_p^{-1}(m_0,\boldsymbol{\Psi}_0)$$

has the random spectral decomposition $\Sigma_k = D_kQ_kD_k^t = \lambda_kD_kAD_k^t$ with random eigenvalues $q_{k1} \geq q_{k2} \geq\ldots\geq q_{kp} \geq 0, Q_k = diag(q_{k1},..,q_{kp})$ and we define $\lambda_k := q_{k1}, \mathbf{A} := diag(1,q_{k2}/q_{k1},\ldots.,q_{kp}/q_{k1})$. In particular, we assume that $\mathbf{A} = diag(1,a_2,\ldots,a_p)$ and $\mathbf{D}_k$ are the shape and direction components of an inverse Wishart random variable $W_p^{-1}(m_0,\boldsymbol{\Psi}_0)$ (for the choice of m_0, $\boldsymbol{\Psi}_\mathbf{0}$ and other priors, see again Section 5). If we assume that $\mathbf{A}$ and $\mathbf{D_k}$ are *a priori independent* (Anderson 1984), the corresponding Gibbs sampler step is to simulate $a_j|\mathbf{D}_1,\ldots,\mathbf{D}_K,\lambda,\nu$, for $j = 1,\ldots,p$, independently from the inverse gamma distribution

$$IG\Big(\tfrac{1}{2}(n+K(m_0+p)-1),\tfrac{1}{2}\big\{\textstyle\sum_k \lambda^{-1}\mathbf{D}_k^t\big((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+\mathbf{W}_k+\boldsymbol{\Psi}_0\big)\mathbf{D}_k\big\}_{jj}\Big)$$

Moreover, the $\mathbf{D}_k$'s are the principal direction vectors from the following inverse Wishart distribution

$$W_p^{-1}\big(n_k+m_0,\boldsymbol{\Psi}_0+\mathbf{W}_k+\tfrac{n_k\tau_k}{n_k+\tau_k}(\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\big).$$

(b) Model $[\boldsymbol{\lambda_kD_kAD_k^t}]$.

Again, $\mathbf{D}_k$ and $\mathbf{A}$ considered here are unknown. In addition, the k different groups are allowed to have different volumes λ_k. The prior distribution of λ_k is assumed to be an inverse gamma distribution

$\lambda_k \sim IG(m_k/2,s_k/2)$ independently for $k = 1,\ldots,K$, and the corresponding Gibbs sampler step 3 is to simulate $a_j|\mathbf{D}_1,\ldots,\mathbf{D}_K,\lambda_1,\ldots,\lambda_K,\boldsymbol{\nu}$, for $j = 1,\ldots,p$, independently from

$$IG\Big(\tfrac{1}{2}(n+K(m_0+p)-1),\tfrac{1}{2}\big\{\textstyle\sum_k \lambda_k^{-1}\mathbf{D}_k^t\big((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\frac{n_k\tau_k}{n_k+\tau_k}+\mathbf{W}_k+\boldsymbol{\Psi}_0\big)\mathbf{D}_k\big\}_{jj}\Big)$$

$\lambda_k|\mathbf{A},\mathbf{D}_1,\ldots,\mathbf{D}_K,\nu$, for $k = 1,\ldots,K$, independently from

$$IG\big(\tfrac{1}{2}(m_k+n_kp),\tfrac{1}{2}\big\{s_k+tr[(\mathbf{D}_k\mathbf{A}^{-1}\mathbf{D}_k^t)\big((\bar{\mathbf{x}}_k-\xi_k)(\bar{\mathbf{x}}_k-\xi_k)^t\tfrac{n_k\tau_k}{n_k+\tau_k}+\mathbf{W}_k+\boldsymbol{\Psi}_0\big)]\big\}\big)$$

The $\mathbf{D}_k$'s are simulated in the same way as in model (a).

(c) General model $[\boldsymbol{\lambda_kD_kA_kD_k^t}]$

This is the standard Gaussian mixture model considered by Lavine and West

(1992). In this case, there is no need to use the eigenvalue decomposition of $\mathbf{\Sigma}_k$. The prior distributions on $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ are assumed:

$$\boldsymbol{\mu}_k|\mathbf{\Sigma}_k \sim N_p(\boldsymbol{\xi}_k, \mathbf{\Sigma}_k/\tau_k), \quad \text{and} \quad \mathbf{\Sigma}_k \sim W_p^{-1}(m_k, \mathbf{\Psi}_k), \quad (k = 1,\ldots,K),$$

and the corresponding Gibbs sampler step 3 is to simulate, $\boldsymbol{\mu}_k|\mathbf{\Sigma}_k,\boldsymbol{\nu}$, for $k = 1,\ldots,K$, independently from

$$N_p(\bar{\boldsymbol{\xi}}_k, \mathbf{\Sigma}_k/(\tau_k + n_k))$$

where $\bar{\boldsymbol{\xi}}_k = (n_k\bar{\mathbf{x}}_k + \tau_k\boldsymbol{\xi}_k)/(n_k + \tau_k)$, and $\mathbf{\Sigma}_k|\boldsymbol{\nu}$, for $k = 1,\ldots,K$, from

$$W_p^{-1}\left(n_k + m_k, \mathbf{\Psi}_k + \mathbf{W}_k + \frac{n_k\tau_k}{n_k + \tau_k}(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{x}}_k - \boldsymbol{\xi}_k)^t\right).$$

3.2.1.1 Model selection

So far we described the models of interest and working in a mixture-model framework, we will use the approximate Bayes factors to compare the models. For a review of Bayes factors, their calculation and their interpretation, see Kass and Raftery (1995). Here, we have to select not only the parametrization of the model but also the number of clusters K.

For simultaneously choosing between two models $\mathrm{M}_1, \mathrm{M}_2$ and deciding on the number of groups, we compute the approximate Bayes factor

$$BF_{1,2} = p(\mathbf{x}|M_2)/p(\mathbf{x}|M_1)$$

with

$$p(\mathbf{x}|M_h) = \int p(\mathbf{x}|\boldsymbol{\theta}_h)p(\boldsymbol{\theta}_h|M_h)d\boldsymbol{\theta}_h,$$

where $\boldsymbol{\theta}_h$ is the vector of parameters under the model M_h, and $p(\boldsymbol{\theta}_h|M_h)$ is its prior density ($h = 1,2$). The quantity defined in (10) is called the *integrated likelihood* of model M_h. Bayesian model selection is based on Bayes factors, whose key ingredient is the integrated likelihood of a model. By convention, $\log(\mathrm{BF}_{12}) < 2$ represents weak evidence for the model M_2, differences between 2 and 6 represent positive evidence, differences between 6 to 10 represent strong evidence, and differences > 10 represents very strong evidence (Jeffreys 1961). We approximate the integrated likelihood from the Gibbs sampler output using the *Laplace-Metropolis estimator* (Raftery 1996), which is very simple to calculate and was shown to give sufficiently accurate results by Lewis and Raftery (1997) and Bensmail et al. (1997). In the sequel, the word ”model” refers to a combination of one of the models in Table 1 with a particular number of clusters K. Using the Laplace-Metropolis estimator, the Bayes factor becomes

$$BF_{1,2} = \frac{p(\mathbf{x}|M_2)}{p(\mathbf{x}|M_1)} = \frac{|\mathbf{\Psi}^{(2)}|^{1/2}p(\mathbf{x}|\tilde{\boldsymbol{\theta}}^{(2)})p(\tilde{\boldsymbol{\theta}}^{(2)})}{|\mathbf{\Psi}^{(1)}|^{1/2}p(\mathbf{x}|\tilde{\boldsymbol{\theta}}^{(1)})p(\tilde{\boldsymbol{\theta}}^{(1)})},$$

where $\tilde{\boldsymbol{\theta}}^{(h)}, (h = 1, 2)$ is the posterior mode of $\boldsymbol{\theta}^{(h)}$, denoting the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the model M_h, and $\boldsymbol{\Psi}^{(h)}$ is minus the inverse Hessian of $g(\boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ under the model h, evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}^{(h)}$. The Laplace method requires us to know the posterior mode, $\tilde{\boldsymbol{\theta}}$, and $|\boldsymbol{\Psi}|$. The Laplace-Metropolis estimator estimates these parameters from the Gibbs sampler output $\tilde{\pi}_k, \tilde{\boldsymbol{\mu}}_k$ and $\tilde{\boldsymbol{\Sigma}}_k$. The likelihood at the approximate posterior mode is

$$\prod_{i=1}^{n} \sum_{k=1}^{K} \tilde{\pi}_k f_k(\mathbf{x}_i | \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)$$

which is then substituted into equation of the Bayes factor.

For choosing the appropriate model, we calculate the Bayes factor for each pair of different combinations for a number of different clusters with the number of the components varying from $1, \ldots, M$ for all models. This procedure is exemplified in Section 5.

3.2.2 Maximum likelihood cluster analysis: mclust

We used the function EMclust to initialize the Bayesian Clustering algorithm. The iterative algorithm also has the option of clustering the data using stand-alone 'Mclust' clustering.

MCLUST provides two functions, Mclust and EMclust, for cluster analysis combining hierarchical clustering, EM (Expectation Maximization), and BIC.

The EM methods provided are the iterative EM (Expectation-Maximization) methods for maximum likelihood clustering with parameterized Gaussian mixture models. In this application, an iteration of EM consists of an 'E'-step, which computes a matrix z such that zik is an estimate of the conditional probability that observation i belongs to group k given the current parameter estimates, and an 'M-step', which computes maximum likelihood parameter estimates given z. In the limit, the parameters usually converge to the maximum likelihood values for the Gaussian mixture model.

$$\prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \Phi_k(x_i | \mu_k, \Sigma_k)$$

where τ_k is the mixing proportions and n is the number of observations in the data. Here G is the number of groups, which is fixed in the EM algorithm. The parameterizations of Σ_k are currently available for EM in MCLUST are listed in Table 1.

In both functions, hierarchical clustering is used to initialize EM for various parameterizations of the Gaussian model. Mclust is intended to be a simplified function for one-step model-based clustering, with reasonable defaults. EMclust has more options and more flexibility, although it may be more complicated to use. EMclust has more flexibility than Mclust for clustering, by allowing choice of hierarchical clustering (which need not be model-based) as input to initialize EM and choice of models for EM. Users can obtain parameters and clustering results through summary for any model or number of mixture components specified, rather than just the maximum-BIC model as in Mclust. The input to EMclust includes the data, a list of models to apply in the EM phase, the desired numbers of groups to consider. Default starting values for EM are obtained from applying the function hc for model-based hierarchical clustering to the data using the unconstrained model VVV. EMclust returns the BIC values for all of the chosen models and number of clusters, together with auxiliary information that is used by the corresponding summary method for recovering parameter values.

3.3 Handling Noises In Cluster Analysis

Mahalanobis distance calculation is used to detect the outliers in the algorithm. This distance was first introduced by P.C Mahalanobis in the year 1936.

For a p-dimensional multivariate sample $x_i(i = 1,2\ldots n)$, the Mahalanobis distance is defined by

$$D_i = ((x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k))^{\frac{1}{2}} \text{ for } i = 1,2\ldots n$$

where μ_k is the estimated multivariate location and Σ_k is the estimated covariance matrix.

It becomes Euclidean distance if the covariance matrix is an identity matrix. So this distance computation has an advantage of considering the covariance matrix, compared with other classical statistical approaches.

These distances are chi-square distributed with degress of freedom equal to the dimensionality of the data. Multivariate outliers are the distances which exceed the quantile (99.5%) of the chi-square distribution.

3.4 Iterative Clustering Algorithm

The software is written in R with Rggobi as the visualization tool. This tool enables the user to observe the points in two dimensional spinning plot and cluster the data points visually looking at them. Clustering can be done using the different colors and glyphs provided by the tool.

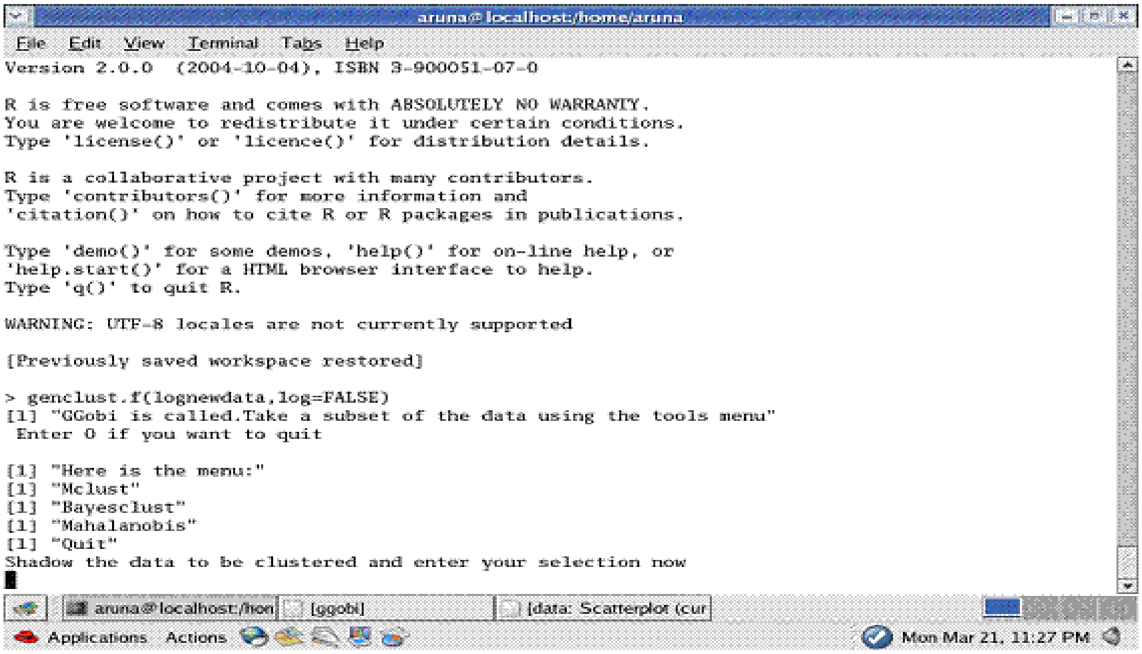
Let $X = \{X1, X2, X3 \ldots \ldots XN\}$ denote the entire data set, where 'N' is significantly very large. Consider the following iterative algorithm:

The software requires to be given the entire data set, and if the log of the dataset to be clustered is required, as an input.

The software prompts the user to enter any information required for the algorithm to continue. One can quit at any time during the execution of the program.

1. Choose a random sample 's' of size 'n' where n<<<N from *X*. This is done from the visualization software – Rggobi.
2. The user needs to shadow the data points that are to be clustered. All the data points, which comprise the sample, can be shadowed to cluster the entire sample at once. While using the Mahalanobis distance measurement also, we need to shadow the data which we think can be potential outliers.
3. Cluster the observations using any appropriate clustering algorithm. Right now, the available algorithms [Figure 3] are simple visual clustering using the visualization software, Mclust and Bayesclust. The visual clustering of the data points will help the user to pertain any details about the grouping of the data points, that the user knows.
4. Mclust uses the Emclust function to get the different clusters of the sample. We need to provide with the number of clusters we anticipate for the shadowed data to be clustered. Different clusters are visually shown with different colors in Rggobi's scatterplot.
5. Bayesian Clustering starts with the input of the sample, number of clusters, type of model and number of iterations. This clustering is an iterative process by itself and the initial clusters are given from the Mclust algorithm. The final clusters from Bayesian clustering are shown on the Rggobi's scatterplot.
6. Repeat steps 2-5 until the user finds that the clusters are repeatedly giving the same number of observations of the sample.

If there are clusters, which are not convincing visually, or statistically the clustering process is overfitting the sample data, we can suspect that there are outliers for the particular clusters. These outliers can be detected using the Mahalanobis distance criterion.



```
Version 2.0.0  (2004-10-04), ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

WARNING: UTF-8 locales are not currently supported

[Previously saved workspace restored]

> genclust.f(lognewdata,log=FALSE)
[1] "GGobi is called.Take a subset of the data using the tools menu"
 Enter 0 if you want to quit

[1] "Here is the menu:"
[1] "Mclust"
[1] "Bayesclust"
[1] "Mahalanobis"
[1] "Quit"
Shadow the data to be clustered and enter your selection now
```

Figure 3: Menu of the iterative algorithm

This distance is almost equivalent to the Euclidean measurement except that it takes the covariance of the cluster also into account. Right now the outliers are depicted with the last color present in the color scheme and the points, which belong to the clusters, are depicted using the same colors as their respective clusters.

Mahalanobis also can be repeated until the software provides with constant number of outliers. Obviously, the outliers get reduced as the process runs iteratively, but there may be certain points which are distinct outliers. So, these outliers by themselves make up another cluster. Again any of the clustering algorithms can be used to cluster these outliers. Now, once the sample is clustered with the outliers and the original clusters, all the remaining data can be added to the process. This can be entirely performed using the Rggobi software. Mahalanobis distance is used to assign the remaining data points to the sample clusters. There may be many outliers here, which might not be captured in the sample data. Small portions of the data points, which are detected as potential outliers can be clustered again by using any of the three clustering algorithms.

CHAPTER 4
RESULTS

4.1 Proteomics Data

Functional data transformation reduces the dimensionality of the spectra. The spectral data were collected from proteomics analysis of a total number of serum samples ($n = 68$) including healthy or normal ($n_1 = 37$), ATL ($n_2 = 20$) and HAM ($n_3 = 11$) patients. The data set is represented by a $n \times p$ matrix **X,** where $p = 25,196$ is the number of variables (peaks) measured on each sample and $n = 68$ is the number of samples (patients). Any clustering algorithm on data ($68 \times 25,196$) will fail because of the singularity of the covariance matrix ($n < p$) and it will have difficult manipulating matrices with 68 rows and 25,196 columns, which has 1. 7133 $\times 10^6$ elements. This problem would not be raised for Heuristic-based (i.e. pairwise similarity based) clustering algorithms.

To reduce the dimensionality of the spectral data, we applied FDA by fitting a P-spline curve $\hat{\boldsymbol{\mu}}_i(t)$ to each sample $\mathbf{y}_i$. P-splines satisfy a penalized residual sum of squares criterion, where the penalty involves a specified degree of derivation for $\boldsymbol{\mu}_i(t)$. For example, cubic splines functions are P-splines of second order, penalizing the second derivative of $\boldsymbol{\mu}_i(t)$. P-splines curves of order 3, penalize the third derivative of $\boldsymbol{\mu}_i(t)$. P-splines curves of order 4, lead to an estimate of $\boldsymbol{\mu}_i(t)$ with continuous first and second derivative. We choose here to fit a P-spline curve of order 4. The fitting step is performed by fixing the number of degrees of freedom that are implicit in the smoothing procedure [23].

The next step performed on the smoothed curves is to find the landmarks or indices T_i. We collected the first derivative of $\hat{\boldsymbol{\mu}}_i(t)$, say $\hat{\boldsymbol{\mu}}_i^{'}(t)$ using a smoothing Pspline function available in *R*. Those derivatives are crucial at determining the cut-off points or indices of $\boldsymbol{\mu}_i(t)$. We performed this step by computing an approximate 95% pointwise confidence interval for the first derivative of $\boldsymbol{\mu}_i(t)$ [24]. When the lower limit of this interval is positive, we have the confidence that $\boldsymbol{\mu}_i(t)$ will be increasing. When the upper limit of this interval is negative, we have the confidence that $\boldsymbol{\mu}_i(t)$ will be decreasing. Inside the interval, when the derivative changes from negative to positive, we have an optimal value which is a minimum. When the derivative changes from positive to negative, we have an optimal value which is a maximum. The maximum is set, for convenience, as the largest value of $\hat{\boldsymbol{\mu}}_i^{'}(t)$ in that interval. In this study, we restricted the choice of indices to maximal values.

Let $\Lambda_i = \left\{\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{m_i}^{(i)}\right\}$ be the collection of the estimated points where the curve $\boldsymbol{\mu}_i(t)$ has a local maximum and let m_i be the number of maximal per observation or per sample (i). Consequently, dissimilarity measure is calculated to derive the dissimilarity matrices of size ($n \times n$) for all samples using the maximum values.

4.1.1 Clustering spectral data using Functional Data Analysis

The application of functional data transformation led to the reduction of the dimensionality of the spectra to half. The size of mass indices become $12,598$. To cluster the reduced data, we calculated the three dissimilarity matrices $M_{\delta_C}, M_{\delta_B}$ and $M_{\delta_{HZ}}$. It appears that an unusual sample (patient 11) hides a possible pattern that we are trying to discover. Figure 4 shows a clustering dendogram of the data using Diana approach. Map and Clara gave the same results. This suggests that sample 11 would be important for further investigation.

When we removed observation 11, we detected a fewer fuzzy patterns with δ_C [Figure 5], δ_{HZ} [Figure 6] and δ_B [Figure 7]. To be more specific, we investigated clusters proposed by δ_C and δ_{HZ}. A large number of of clusters were proposed by both approach (about 10 clusters). This strange result might be caused by the monotonocity assumption when using δ_{HZ} or the lost of informations when using δ_C .

For δ_B, we provided the dendogram of the data using Diana approach [Figure 8]. Three clusters were apparent. Clara also showed the same result [Figure 9], one well separated clusters and two overlapped ones. For δ_{HZ} and δ_C, no structure was apparent, which confirms the limitations of both dissimilarities as explained before.

To check the performance of our method, we calculated the confusion matrix between the predicted clusters and the clinical clusters [Table 2]. We find that 3 patients out of 11 were misclassified for Cluster 1 (HAM), 6 out of 20 were misclassified for Cluster 2 (ATL) and 3 out of 37 were misclassified for Cluster 3 (Normal). HAM and ATL shared the majority of the misclassified observations which makes sense since both groups gathers patients with a disease caused by the same retrospective virus. The error rate of misclassification for both clusters (HAM and ATL) is about 20%. For normal patient, the error rate of misclassification is about 8%. The total rate of misclassification is about 16%.

When we used Clara-based hierarchical cluster algorithm with δ_B, the classification result has dramatically been improved [Table 3]. The error rate of misclassification is reduced to 7%. The error rate of misclassification between HAM and ATL is about 9%, while 5% of normal patients was misclassified.

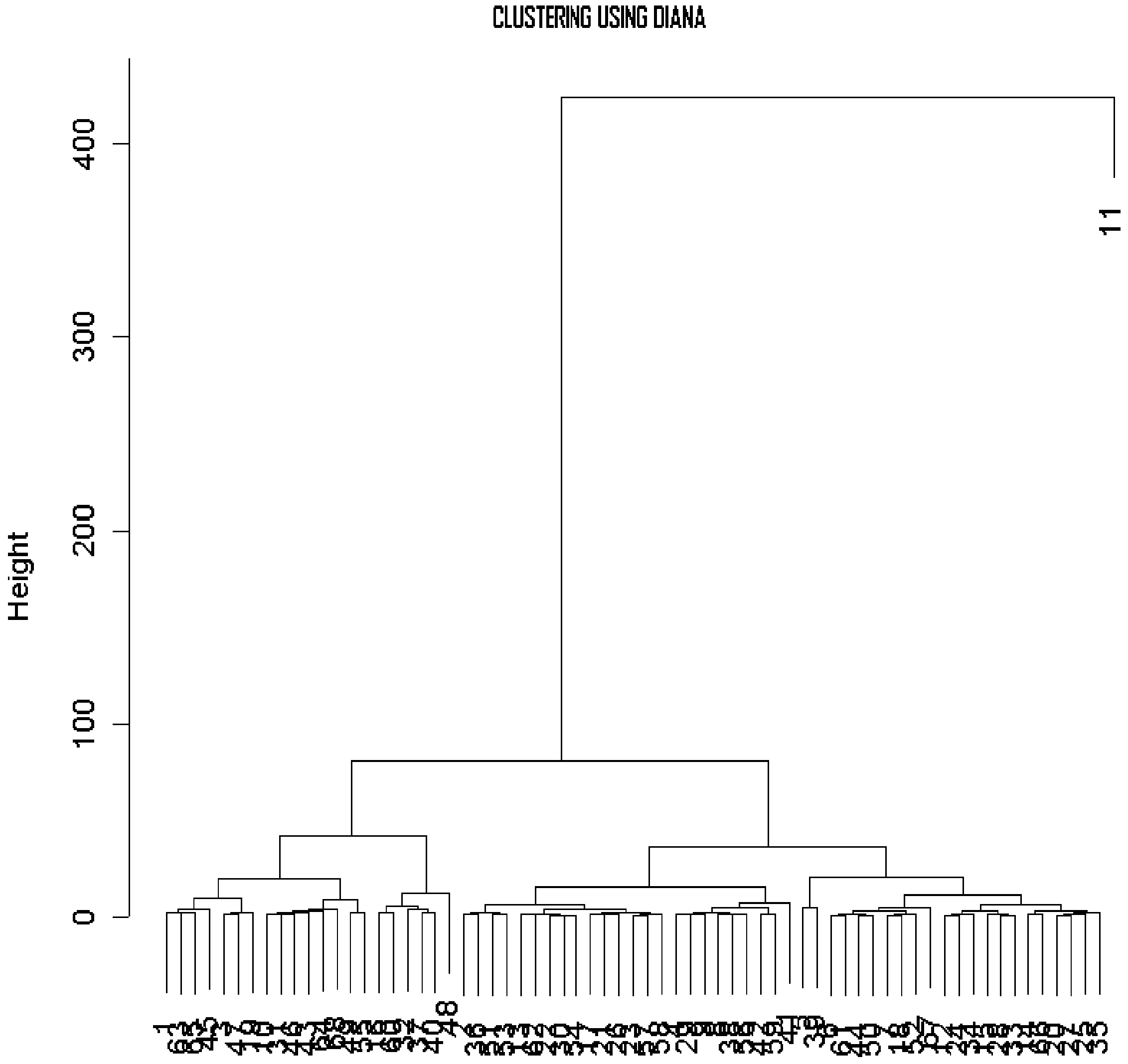


Figure 4: Clustering proteomics data with diana

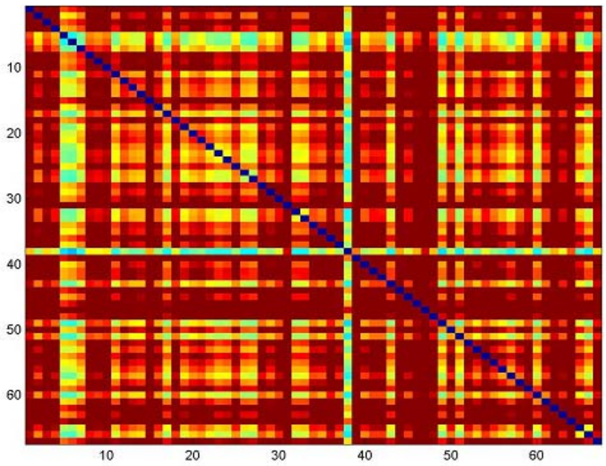


Figure 5: Pattern recognition using δ_C

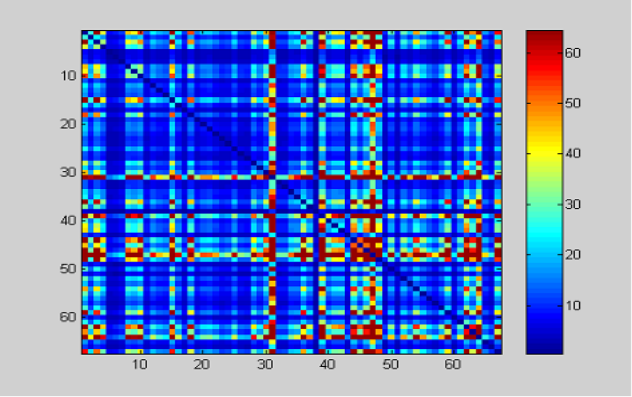


Figure 6: Pattern recognition using δ_{HZ}

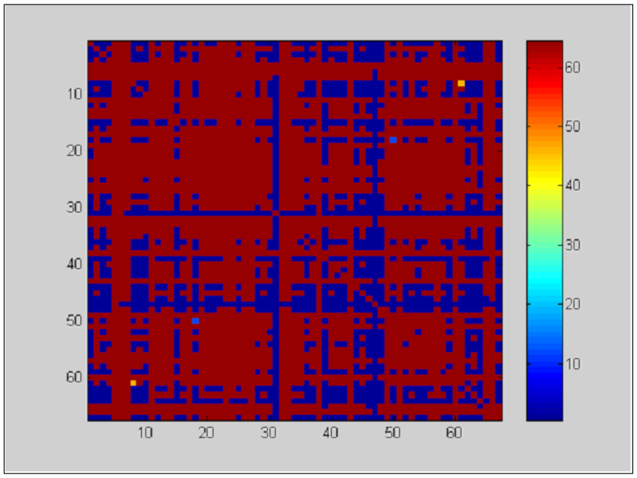


Figure 7: Pattern recognition using δ_B

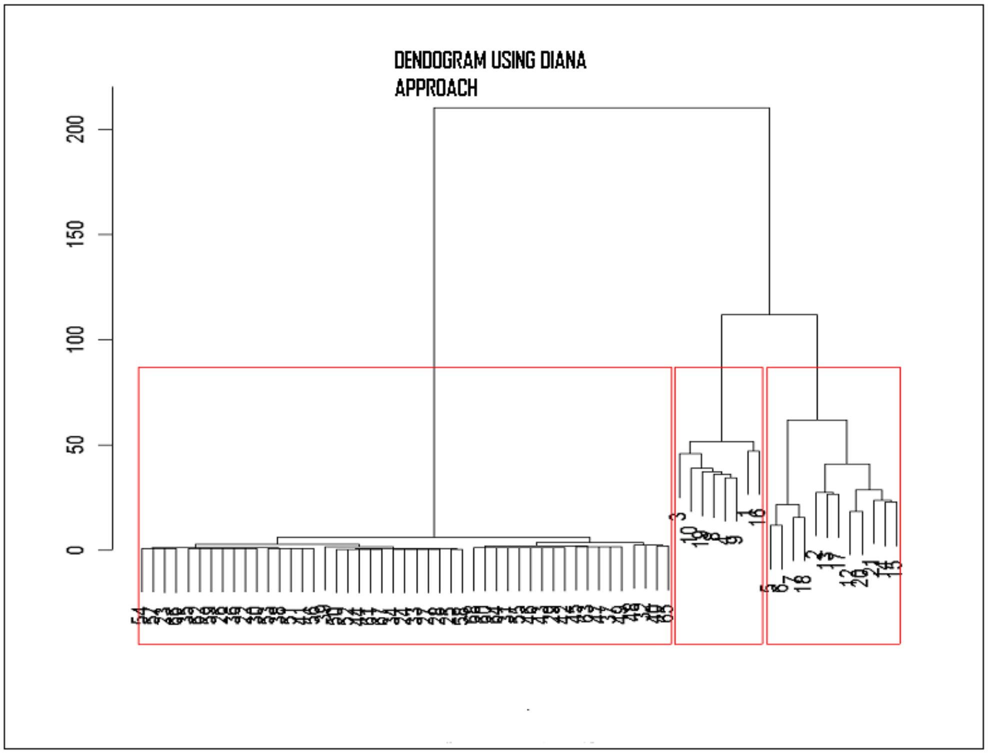


Figure 8: Dendogram of the δ_B –dissimilarity approach with diana.

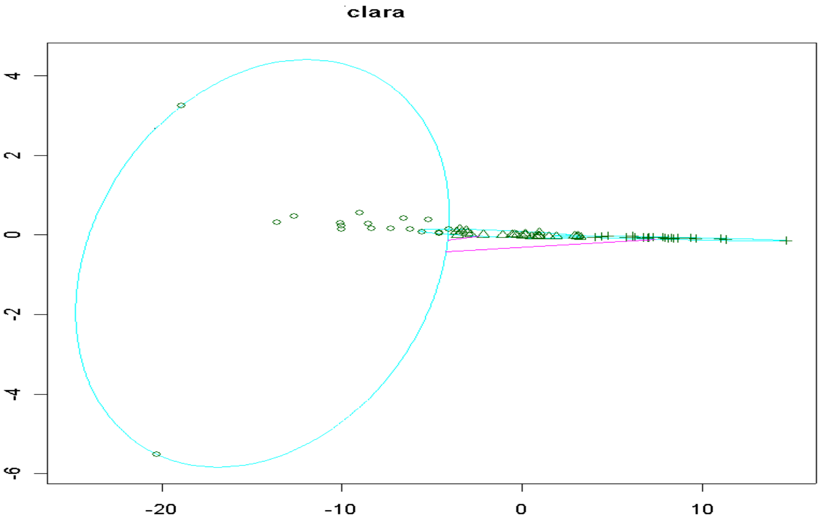


Figure 9: The δ_B –dissimilarity approach with clara

Table 2: Confusion matrix to show the performance of δ_B using diana.

		Predicted			Total
		HAM	ATL	Normal	
	HAM	8	3	0	11
Clinical	ATL	5	14	1	20
	Normal	1	2	34	37
Total %		0.73	0.70	0.92	0.84

Table 3: Confusion matrix to show the performance of δ_B using clara.

		Predicted			Total
		HAM	ATL	Normal	
	HAM	10	0	0	11
Clinical	ATL	2	18	0	20
	Normal	1	1	35	37
Total %		0.91	0.90	0.95	0.93

This results shows that a hierarchical δ_B –dissimilarity algorithm based on minimizing the dissimilarity of observations to their closest medoid performs better than a divisive hierarchical clustering algorithm based on δ_B.

4.2 Spatial Data

4.2.1 Example 1: Simulated data

The first example is the simulated data set with number of observation $n = 286,320$ and the number of variables $p = 4$.

The log of the data at first looks as in the Figure 10.The initial step of the software prompts the user to enter as what one needs to do following the menu choices. We can also quit using the 'quit' option. Following the second step of the process, a random sample of the data is taken, which can be depicted from the Rggobi tool.

The tool of Rggobi provides with an feature of three – dimensional rotating plot. So any of the clusters, which are not seen, from one perspective are discovered in another. Figure 11 is the rotating version of the sample data

Thus, looking at the data, we can conclude that there are some clusters that can be manually clustered. Hense, I used different colors from the color palette of the Rggobi brushing tool. The clusters in the initial aqua color are the clusters, which are clustered in the following stages. So these are to shadowed [Figure 12]. The shadowed data is clustered into six different clusters in different colors, with the option of Bayesian clustering, thus making the total of ten clusters [Figure 13].

But the clusters right in the middle of the plot, may be just outliers to the other clusters, so Mahalanobis option is chosen to find the outliers. The output from R when Mahalanobis is used.All the Cluster means and covariances are calculated first, and the Mahalanobis module function gives the number of outliers present. Right now, there are 276 outliers out of 277 observations.The Mahalanobis o/p from the Rggobi is shown in Figure 14.The outliers are given the last color which is brown in the present palette.

Now, the entire data is added to the sample using the Rggobi tool. Then Mahalanobis is used to assign the entire dataset to the sample and is shown in [Figure 15].

The clusters in brown are the distinct outliers, which need to clustered again. So the final clusters can be seen in Figure 16.

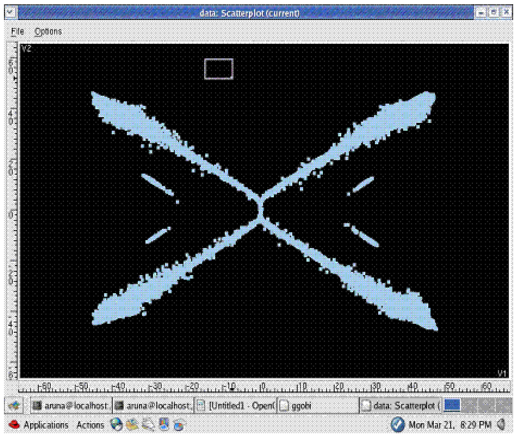

data: Scatterplot (current)
File Options
V2
V1
-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60
aruna@localhost: aruna@localhost: [Untitled1 - Open(ggobi data: Scatterplot (
Applications Actions
Mon Mar 21, 8:29 PM

Figure 10: Log of the data

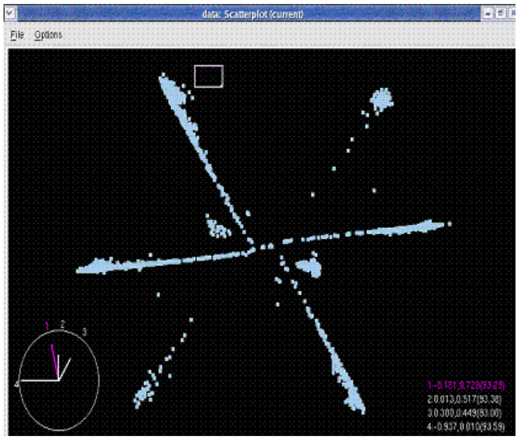


Figure 11: Sample of the data

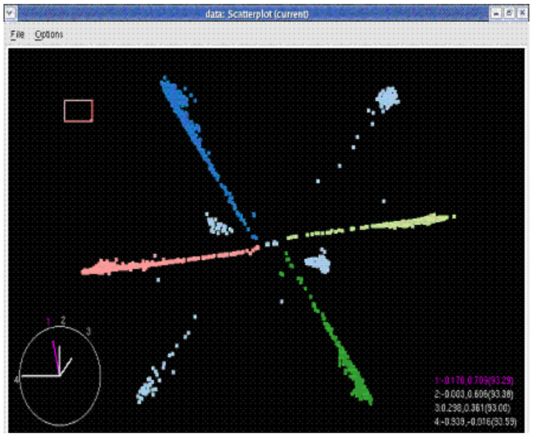


Figure 12: Visual clusters

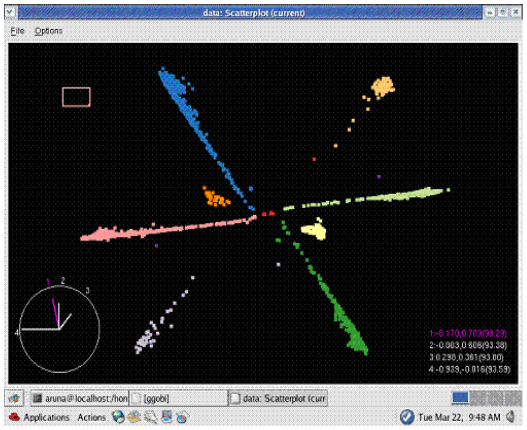


Figure 13: Clusters without outliers

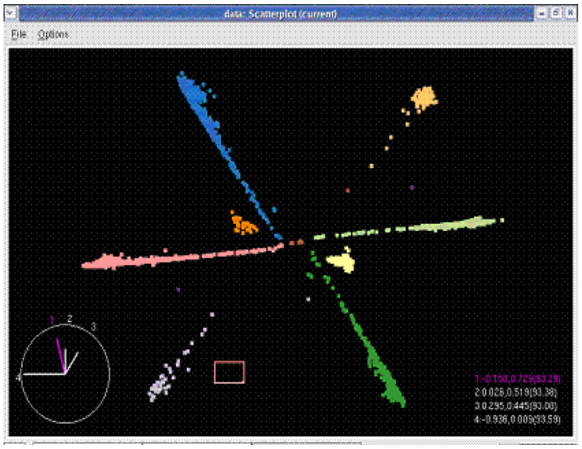


Figure 14: Clusters with outliers

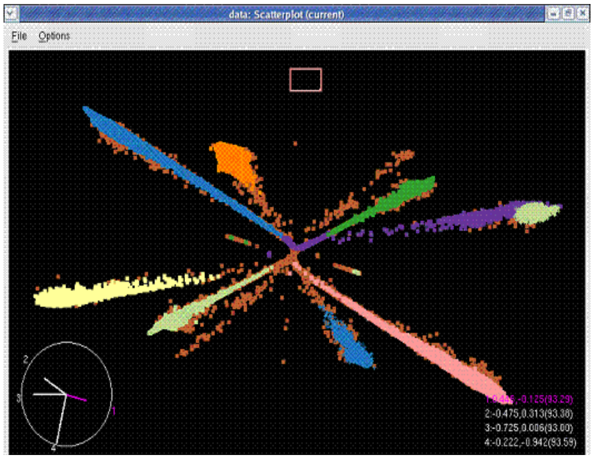
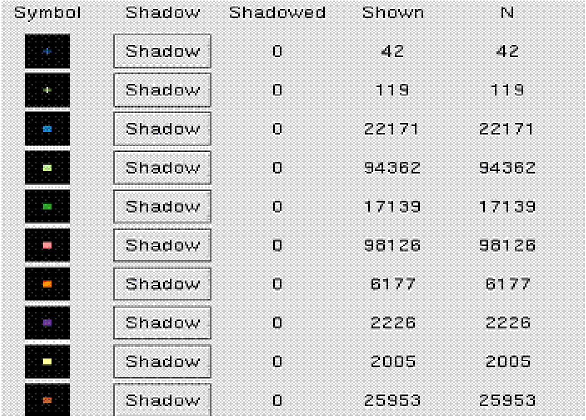


Figure 15: Final clusters



Symbol	Shadow	Shadowed	Shown	N
+	Shadow	0	42	42
+	Shadow	0	119	119
🟦	Shadow	0	22171	22171
🟩	Shadow	0	94362	94362
🟩	Shadow	0	17139	17139
🟥	Shadow	0	98126	98126
🟧	Shadow	0	6177	6177
🟪	Shadow	0	2226	2226
🟨	Shadow	0	2005	2005
🟫	Shadow	0	25953	25953

Figure 16: Clusters of the data

4.2.2 Example 2: Real data

Stars are generally classified into two main groups called Population I and Population II. The stars of the two populations are very similar. They all use fuse elements to generate energy in the same way and they follow the same evolutionary sequence. However, there are important differences in the characteristics which distinguish the two groups; one of the main differences lies in the metal content of the stars in each group. Astronomers regard all elements other than Hydrogen and Helium as a ”metals” because metals were not originally part of the makeup of the universe, but were manufactured inside the cores of heavy stars and dispersed throughout the Galaxy by stellar winds, stellar flashes, and supernova explosions.

Population I stars are relatively metal rich stars; they contain about 2-3 percent metals. Because these stars have high metal content, astronomers know that they are not first-generation stars, but rather were formed from the material fused in earlier generations of stars. Population I stars inhabit the disk of the Galaxy. They travel in approximately circular orbits about the center of the Galaxy and they generally remain in the plane of the Galaxy as they orbit. (The older Population I stars are found farther out of the plane than the younger stars.) Population I stars are relatively young stars because they have formed within the last few billion years. Extreme Population I stars (the most metal rich stars) are found only in the spiral arms; these are the youngest stars. Intermediate Population I stars, like our Sun, are located throughout the disk. They are slightly less metal rich. As the galaxy ages, subsequent generations of stars will become increasingly metal rich because more and more heavy elements will be fused inside massive stars, and their subsequent explosions will further enrich the galactic interstellar medium.

By contrast, Population II stars are metal poor; they contain about 0.1 percent metals. They are found in the spherical components of the Galaxy, the halo and the bulge. They have randomly tipped, elliptical orbits which can plunge through the disk of the Galaxy and which take some of them (the halo stars) to large distances from the center. They are relatively old stars, with ages ranging from 2 - 14 billion years. Extreme Population II stars (the most metal poor) are found in the halo and the globular clusters which orbit the center of the galaxy; these are the oldest stars. Intermediate Population II stars are located in the bulge. They are slightly more metal rich than the extreme Population II stars, but less metal rich than the intermediate Population I stars.

The differences in metalicity and age between the two Populations implies that the Population II stars formed early during the evolution of the Galaxy. At this time, the Galaxy would have contained gas that was nearly pure hydrogen and helium, because few stars would have had enough time to generate heavier elements and disperse these heavier elements into the Galaxy. Consequently, our Galaxy consists of two stellar populations, the disk and the halo. More recently it has been hypothesized that there are in fact three stellar populations: the old (thin) disk, the thick disk, and the halo, distinguished by their spatial distributions, their velocities and their metallicities. These hypotheses have different implications for theories of formation of the Galaxy. Some of the evidence for deciding whether there are two or three populations is given in Figure 17, which shows radial and rotational velocities for $n = 2370$ stars from Soubiran (1993). In Celeux and Govaert paper (1995), authors used the spherical models $[\lambda I]$ and $[\lambda_k I]$. The spherical model with the same volume was not able to clearly distinguish a dense cluster (Halo or population II). The spherical model with different volume on the contrary, shows its ability to find clusters with different sizes.

Here we can show that we can do better since we do not know if the population is spherical or ellipsoidal. By considering all the models available in Table 1, we will let the data speak out. Figure 18 presents a random sample δ_0 randomly selected from the Galaxy data set D. Figure 19 and Table 4 show that the model $[\lambda_k \Sigma]$ (which means that clusters have different sizes and same shape) is preferred and that there is a strong evidence for three groups as against two.

The balance of astronomical opinion has also tilted towards this conclusion, based on more information than just velocities used here. Other information includes star positions and metallicities (Soubiran 1993). We reached this conclusion with the present method (Bayesian finite mixture model) using only a relatively small part of the total available information. No outliers was detected by our method.

The posterior means of the parameters for the preferred model are: $\lambda_1 = 1$, $\lambda_2 = 11.2$, $\lambda_3 = 2$; $\mu_1 = (-10, -10)$, $\mu_2 = (2.5, -100)$, $\mu_3 = (16, -38)$;

$$\Sigma = \begin{pmatrix} 1045 & -22 \\ -22 & 550 \end{pmatrix}$$

Table 4: Bayes factors for different models

No. groups	$[\lambda I]$	$[\lambda_k I]$	$[\Sigma]$	$[\lambda_k \Sigma]$
1	2639	2641	2642	2641
2	2656	2628	2642	2621
3	2712	2601	2716	**2566**
4	2844	2624	2716	2670
5	2811	2821	2765	2713
6	2823	2822	2851	2810

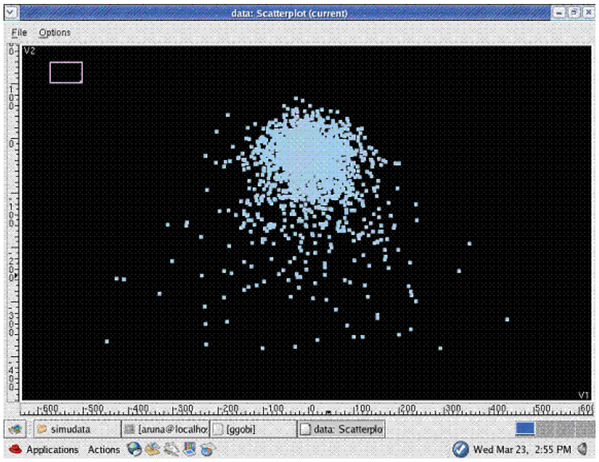


Figure 17: Real data

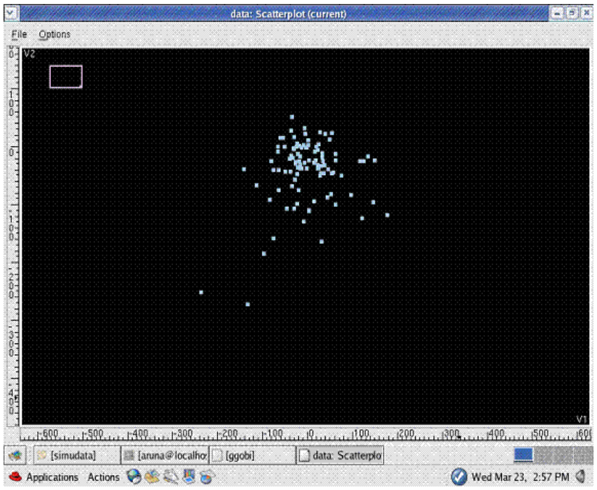


Figure 18: Sample data

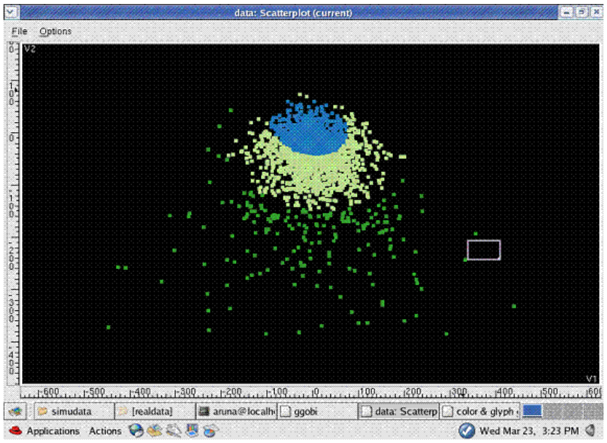


Figure 19: Final Bayesian clusters

CHAPTER 5
DISCUSSION

Cancer biomarkers can be used to screen asymptomatic individuals in the population, assist diagnosis in suspected cases, predict prognosis and response to specific treatments, and monitor patients after primary therapy. The introduction of new technologies to the proteome analysis field such as mass spectrometry have sparked new interest in cancer biomarkers allowing for more effective diagnosis of cancer by using complex proteomic patterns or for better classification of cancers, based on molecular signatures, respectively. These technologies provide a wealth of information and rapidly generate large quantities of data.

Processing the large amounts of data will lead to useful predictive mathematical descriptions of biological systems which will permit rapid identification of novel therapeutic targets and diseases biomarkers. Clustering and analyzing Proteomics data has proven to be a challenging task.

Proteomics data are provided usually as curves or spectra with thousand of peaks. A clustering algorithm based on a matrix of n observations (n samples which is usually small) and p peaks (p variables which is usually a large number) will be unsuccessful. A covariance matrix of size ($n \ll p$) will be singular and any method based on a matrix M ($n \times p$) will not be robust enough and will induce errors. A clustering algorithm based on a well chosen dissimilarity matrix ($n \times n$) is more appropriate and more robust given the relatively moderate size of the matrix.

The application of Euclidean or Mahalanobis distances for instance may not perform well for this proteomics data set, since those distances are usually successfully applied to a typical data with specific expression, spherical or ellipsoidal (Normally distributed data). A new dissimilarity measure has to involve other criteria such as the wealth of data points for each observation and the parallel nature expressed by the proteomics curve (or time series). On the other hand, a robust dissimilarity measure may perform badly on a curve with a large data points or peaks.

Functional smoothing of proteomics expression profiles or spectra has proven to be very helpful. This has allowed us to minimize the number of peaks to retain only the ones that passed the performance of the FDA smoothing. In this study, after using FDA, we succeeded in retaining 50% of the smoothed peaks. The FDA with the dissimilarity measure δ_B shows better performance by comparison to δ_C and δ_{HZ} known to perform well along with FDA on times series data or on monotonic curves.

The two remaining difficulties that naturally arose are (1) to find meaningful peaks that can be used to provide better discrimination between the clusters, and (2) to propose the optimal number of clusters instead of choosing them a priori. The model selection criteria might be useful to answer those questions. In fact, model selection scores use two components for selecting the number of variables and the number of clusters in a given density-based cluster analysis. The first term is the lack of fit generally proportional to the likelihood function. The second term is the penalty term (complexity term). For such proteomics data set, we propose to use the sum of the negative δ_B dissimilarity measure between all the observations to their closest medoids as a lack of fit function. The penalty term might be simple to derive but biased using AIC and BIC, for example, or it can be more difficult to derive if one used a more robust method such as an Information complexity based criterion.

The iterative algorithm being the combination of automated and visual clustering methods performed well in the clustering of simulated data as well as real data. Clusters in the simulated data seem to be convincing visually that all the clusters have been discovered by the algorithm. The algorithm performed well on the real data also. Three clusters have been discovered in the astronomical data. So, there is a stronger evidence of the real data being clustered into three clusters rather than two.

There might be some difficulties in clustering the data, especially massive and gigantic data sets with the itearative software, since RGGobi is used as the visualization software. Sometimes, the Rggobi tool may not perform well with the data sets bigger than a million observations, in which case, visual clustering may become a hassle. Also, the installation of the software in the Linux platform has some problems since the documentation of the RGGobi is not sufficiently clear. But one advantage of RGGobi is the three dimensional spinning plots, which make the multi-scale data clustering more obvious.

Some future work may include working on the clustering of the non-normal or non-gaussian data. Right now, in the bayesian clustering, data is assumed to be normal. Also, empty clusters which result in some of the iterations of the Bayesian clustering method, are now treated by adding some points from the largest cluster to make the algorithm work. Some other technique can be used to treat these empty clusters.Thirdly, better initialization of the clusters can be given to the Bayesian method at the start.

REFERENCES

[1] Aebersold R., Mann M. (2003) Mass spectrometry-based proteomics. Nature 422, 198-207

[2] Steen H, Mann M. (2004) The ABC's (and XYZ's) of peptide sequencing. Nat Rev Mol Cell Biol. 5, 699-711.

[3] Wright GL. (2002) SELDI proteinchip MS: a Platform for biomarker discovery and cancer diagnosis. Expert Rev Mol Diagn. 6, 549-563

[4] Reddy G, and Delmasso EA. (2003) SELDI ProteinChip(R) Array Technology: Protein-based predictive medicine and drug discovery applications. J Biomed Biotechnol 2003, 237-241.

[5] Tang N, Tornatore P, Weinberger SR. (2004) Current developements in SELDI affinity technology. Mass Spectrom Rev. 23, 34-44

[6] Espina V., Mehta AI., Winters ME., Calvert V., Wulfkuhle, J., Petricoin EF 3rd, Liotta LA. (2003) Protein microarrays: molecular profiling technologies for clinical specimens. Proteomics, 3, 2091-2100.

[7] Zhang H, Yan W, Aebersold, R. (2004) Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. Curr Opin Chem Biol. 8, 66-75

[8] Vazquez A, Flammini A., Maritan A, Vespignani A. (2003) Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 21, 697-700.

[9] Bensmail H, Haoudi A (2003) Postgenomics: Proteomics and bioinformatics in cancer research. J Biomed Biotechnol, 2003, 217-230.

[10] Sumorjai RL, Dolenko B, Baumgartner R. (2003) Class prediction and discovery using gene microarray and proteomics mass spectrometry data: curses, caveats, cautions. Bioinformatics 19, 1484-1491.

[11] Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM. (2004) Protein profiling in brain tumors using mass spectrometry: feasibility of a new technique for the analysis of protein expression. Clin Cancer Res 10, 981-987

[12] Ramsay and Silverman (1997) *Functional Data Analysis*, Springer, New York

[13] Ramsay and Silverman (2002) *Applied Functional Data Analysis*. Methods and case studies, Springer, New York.

[14] Haoudi, A. and Semmes, O.J. (2003) The HTLV-1 tax oncoprotein attenuates DNA damage induced G1 arrest and enhances apoptosis in p53 null cells.. *Virology* **305**, 229–239.

[15] Haoudi, A., Daniels, R.C., Wong, E., Kupfer, G. and Semmes, O.J. (2003) Human T-cell Leukemia Virus-I Tax Oncoprotein Functionally Targets a

Subnuclear Complex Involved in Cellular DNA Damage-Response. *J. Biol. Chem.*, **278**, 37736-3774

[16] Adam,B.L., Qu,Y., Davis,J.W., Ward,M.D., Clement,M.A., Cazares,L.H., Semmes,O.J., Schelhammer,P.F., Yasui,Y., Ziding,F. and Wright,G.L. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, **62**, 3609-3614.

[17] Piccolo, D. (1990) A distance measure for classifying ARIMA models, *Journal of Time Series Analysis*, 11, 153-164.

[18] Corduas, M. (2000) La metrica autoregressiva tra modelli ARIMA: una procedura in linguaggio GAUSS, *Quaderni di statistica*, 2, 1-37.

[19] Heckman, N. and Zamar, R. (2000) Comparing the shapes of regression function, *Biometrika*, 87, 135-144.

[20] Cerioli,A., Laurini,F., Corbellini, A. (2003) Functional cluster analysis of financial time series, *Proceedings of the meeting of classification and Data Analysis Group of the Italian Statistical Society (CLADAG* 2003*)*, CLUEB, Bologna,107-110.

[21] Ingrassia S., Cerioli A., and Corbellini A. (2003) Some issues on clustering of functional data, in: Between Data Science and Applied Data Analysis, Schader M., Gaul W. and Vichi M., eds., Springer-Verlag, Berlin, in print.

[22] Kaufman L. and Rousseeuw P. J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[23] Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*, Chapman and Hall, London.

[24] Silverman, B. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting, *Journal of the Royal Statistical Society,* B, 47-52.

[25] Bensmail, H., Semmens, J. and Haoudi, A. (2005). Bayesian Fast-Fourier Transform Based Clustering Method for Proteomics Data. Journal of Bioinformatics. In Press

[26] ANDERS, K.H.(2001), ”Data Mining for Automated GIS Data Collection,” Photogrammetric Week 2001.

[27] ANDERSON, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

[28] BANFIELD, J. D. and RAFTERY, A. E., (1993), ”Model-Based Gaussian and Non Gaussian Clustering”, *Biometrics,* 49, 803-821.

[29] BENSMAIL, H., CELEUX, G., RAFTERY, A. E. and ROBERT, C.

(1997), ”Inference in Model-Based Cluster Analysis”, *Computing and Statistics,* 7, 1-10.

[30] BINDER, D. H. (1978), ”Bayesian Cluster Analysis,” *Biometrika,* 65 (1), 31-38.

[31] BINDER, D. H. (1981), ”Approximations to Bayesian Clustering Rules,” *Biometrika,* 68 (1), 275-285.

[32] BOCK, H. H. (1985), ”On Some Significance Tests in Cluster Analysis,” *Journal of Classification*, 2, 77-108.

[33] BOCK, H. H.(1996), ”Probability Models in Partitional Cluster Analysis,” *Computational Statistics and Data Analysis*, 23, 5-28.

[34] BOZDOGAN, H. (1993), ”Choosing the Number of Component Clusters in the Mixture Model Using a New Informational Complexity Criterion of the Inverse Fisher Information Matrix,” *O. Opitz, B. Lausen, and R. Klar*, Eds., Information and Classification, Springer-Verlag, 40-54.

[35] BUJA, A., COOK, D., SWAYNE, D., LANG, D.T.(2004), ”Using the R-GGobi Link”, R-GGobi Manual

[36] CELEUX, G., and ROBERT, C. (1993), ”Une Histoire de Discretisation (avec Commentaires),” *La Revue de Modulad*, 11, 7-44.

[37] CELEUX, G., and SOROMENHO, G. (1996), “An entropy Criterion for Assessing the Number of Clusters in a Mixture Model”, *Journal of Classification,* 13(1), 1996.

[38] COHN, D., CARUANA, R., McCALLUM, A., ”Semi-supervised Clustering with User Feedback”, AAAI.

[39] DASGUPTA, A., and RAFTERY, A. E. (1998), ”Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering,” *Journal of the American Statistical Association*, 93, 294-302.

[40] DIEBOLT, J., and ROBERT, C. P.(1994), ”Bayesian Estimation of Finite Mixture Distributions,” *Journal of of the Royal Statistical Society, Series B,* 56, 363-375.

[41] EDWARDS, W., LINDMAN, H., and SAVAGE, L. J. (1963), ”Bayesian Statistical Inference for Psychological Research,” *Psychological Review*, 70, 193-242.

[42] ENGELMAN, L., and HARTIGAN, J. A. (1969), ”Percentage Points of a Test for Clusters,” *Journal of the American Statistical Association*, 64, 1647-1648.

[43] FILZMOSER, P, ”A multivariate outlier detection method, ”Technical Report ,Department of Statistics and Probability Theory

[44] FRALEY, C. (1999), ”Algorithms for Model-Based Gaussian Hierarchical Clustering,” *SIAM Journal on Scientific Computing*, 20, 270-281.

[45] FRALEY, C., and RAFTERY, A.(2003), ”MCLUST: Software for model-based clustering, density estimation and discriminant analysis” Technical Report No 415

[46] FRALEY, C., and RAFTERY, A.(1998), ”How Many Clusters? Which Clustering Method? - Answers via Model-Based Cluster Analysis”, *Computer Journal*, 41, 578-588.

[47] GELMAN, A., CARLIN, J. B., STERN, H. S., and RUBIN, D. B. (1995), *Bayesian data analysis*. Chapman and Hall.

[48] GILKS, W. R., RICHARDSON, S., SPIEGELHALTER, D., J. (1996), ”*Markov Chain Monte Carlo Methods in Practice”*. New York: Chapman and Hall.

[49] GREEN, P. G., (1995), ”Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination”, *Biometrika*, 82, 711-732 (1995).

[50] GORDON, A. D. (1999), *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall, 2nd Eds. New York.6

[51] HARTIGAN, J. A. (1975), *Clustering Algorithms*, Wiley, New York.

[52] JEFFREYS, H. (1961) , *Theory of Probability*, Clarendon.

[53] KASS, R. E., and RAFTERY, A. E. (1995), ”Bayes Factors,” *Journal of the American Statistical Association*, 90, 773-795.

[54] KAUFMAN, L., and ROUSSEEUW, P. J. (1990), *Finding Groups in Data*, Wiley, New York.

[55] LANG, D.T.(2004), ”The R-GGobi Interface”, R-GGobi Manual

[56] LANG, D.T., SWAYNE, A.F (2001), ”GGobi meets R: an extensible environment for interactive dynamic data visualization”, DSC 2001 Proceedings of the 2nd InternationalWorkshop on Distributed Statistical Computing

[57] LAVINE, M., and WEST, M. (1992) , ”A Bayesian Method for Classification and Discrimination,” *Canadian Journal of Statistics*, 20, 451-461.

[58] LEROUX, B. G., (1992), ”Consistent Estimation of a Mixing Distribution”, *The Annals of Statistics*, 20, 1350-1360.

[59] LEWIS, S. M., and RAFTERY, A. E. (1997), ”Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator,” *Journal of the American Statistical Association*, 92, 438, 648-655.

[60] MACQUEEN, J. B. (1967), ”Some Methods for Classification and Analysis of Multivariate Observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 281-297.

[61] MAITHRA, R.(2001), ”Clustering Massive Datasets with Applications in Software Metrics and Tomography,” in Technometrics (Vol 43) No 3.

[62] MCLACHLAN, G. (1982), *The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis*, (Vol. 2), Amsterdam: Handbook of Statistics, in P. R. Krishnaiah and L. N. Kanal, 199-208.

[63] MCLACHLAN, G., and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

[64] MCLACHLAN, G., and Peel, D. (2000). *Finite Mixture Models*. New York; Wiley.

[65] MENZEFRICKE, U. (1981), ”Bayesian Clustering of Data Sets,” *Communication in Statistics-Theory and Methods A*., 10, 65-77.

[66] MEULMAN, J. J. (1986), *A Distance Approach to Nonlinear Multivariate Analysis*, DSWO Press, Leiden.

[67] MEULMAN, J. J., ZEPPA, P., BOON, M. E., and RIETVELD, W. J. (1992), ”Prediction of Various Grades of Cervical Preneoplasia and Neoplasia on Plastic Embedded Cytobrush Samples: Discriminant Analysis with Qualitative and Quantitative Predictors,” *Analytical and Quantitative Cytology and Histology*, 14, 60-72.

[68] MILENOVA, B.L. and CAMPOS, M. M.”O-cluster: Scalable Clustering of Large High Dimensional Data Sets”, Oracle Data Mining Technologies.

[69] MUKHERJEE, M., and FEIGELSON, E. D., BABU, G. J., MURTAGH, F., FRALEY, C., and RAFTERY, A. (1998), ”Three Types of Gamma Ray Bursts,” *Astrophysical Journal*, 508, 314-327.

[70] MURTAGH, F., and RAFTERY, A. (1984), ”Fitting Straight Lines to Point Patterns”, *Pattern Recognition*, 17, 479-483.

[71] RAFTERY, A. E. (1996), ”Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models,” *Biometrika*, 83, 251-266.

[72] RAYMOND, T.Ng and HAN, JIAWEI. (1994), ”Efficient and Effective Clustering Methods for Spatial Data Mining,” Proceedings of the 20th VLDB Conference.

[73] REAVEN, G. M., and MILLER, R. G. (1979), ”An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis,” *Diabetologia*, 16, 17-24.

[74] RICHARDSON, S., and GREEN, P. J. (1997), ”On Bayesian Analysis of Mixtures with an Unknown Number of Components,” *Journal of the Royal Statistical Society, Series B,* 59”, 731-792.

[75] ROEDER, K., and WASSERMAN, L. (1997), ”Practical Bayesian Density Estimation Using Mixture of Normals,” *Journal of the American Statistical Association,* 92, 439, 894-902.

[76] SCHWARZ, G. (1978), ”Estimating the Dimension of a Model,” *The Annals of Statsitics*, 6, 461-464.

[77] SCOTT, A. J., and SYMONS, M. J.(1971), ”Clustering Methods Based on Likelihood Ratio Criteria,” *Biometrics*, 27, 387-397.

[78] SYMONS, M (1981), ”Clustering Criteria and Multivariate Normal Mixtures,” *Biometrics*, 37, 35-43.

[79] TANNER, M., and WONG, W. (1987), ”The Calculation of Posterior Distributions by Data Augmentation (with Discussion),” *Journal of the American Statistical Association*, 82, 528-550.

[80] TIERNEY, L. (1994), ”Markov Chains for Exploring Posterior Distributions,” *Annals of Statistics*, 22, 1701-1762.

[81] WEI, G. C. G, and TANNER M. A. (1990), ”A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithms”, *Journal of the American Statistical Association*, 85, 699-704.

[82] WOLFE, J. H. (1978), ”Comparative Cluster Analysis of Patterns of Vocational Interest,” *Multivariate Behavioral Research* ,13, 33-44.

VITA

Aruna Buddana was born in Machilipatnam, India. She finished her schooling in Machilipatnam, India. She graduated from Lady Ampthil Jr College, Machilipatnam,after her post-school education. She then went to the J.N.T.U College of Engineering, Kakinada, and obtained her Bachelor of Technology degree in Electronics and Communications Engineering in 2001. She then joined the Ball State University, Muncie to pursue her graduate studies in Computer Science. After graduating from Ball State, she extended her graduate studies at The University of Tennessee, Knoxville in the department of statistics. Subsequently she has been doing her research under the able guidance of Prof. Halima Bensmail. She plans to graduate with a Master's degree in Statistics in May 2005.