A Novel approach for Clustering Proteomics Data using Bayesian Fast Fourier Transform

Halima Bensmail¹*, Jennifer Golek¹, Michelle M. Moody², O. John Semmes² and Abdelali Haoudi² *
¹University of Tennessee, Department of Statistics
334 Stokely Management Building, Knoxville, TN 37996-0532
²Eastern Virginia Medical School
Department of Microbiology and Molecular Cell Biology
Norfolk, VA 23501

* To Whom correspondence should be addressed
ABSTRACT:

Motivation: Bioinformatics clustering tools are useful at all levels of proteomic data analysis. Proteomics studies can provide a wealth of information and rapidly generate large quantities of data from the analysis of biological specimens. The high dimensionality of data generated from these studies requires the development of improved bioinformatics tools for efficient and accurate data analyses. For proteome profiling of a particular system or organism, a number of specialized software tools are needed. Indeed, significant advances in the informatics and software tools necessary to support the analysis and management of these massive amounts of data are needed.

Clustering algorithms based on probabilistic and Bayesian models provide an alternative to heuristic algorithms. The number of clusters (diseased and non diseased groups) is reduced to the choice of the number of components of a mixture of underlying probability. The Bayesian approach is a tool for including information from the data to the analysis. It offers an estimation of the uncertainties of the data and the parameters involved.

Results: We present novel algorithms that can organize, cluster and derive meaningful patterns of expression from large scaled proteomics experiments. We processed raw data using a graphical-based algorithm by transforming it from real space data-expression to a complex space data-expression using discrete Fourier transformation; then we used a thresholding approach to denoise and reduce the length of each spectra. Bayesian clustering was applied to the reconstructed data. In comparison with several other algorithms used in this study including K-means, SOM (Kohonen mapping), and LDA (Linear Discriminant Analysis), the Bayesian-Fourier model-based approach displayed superior performance consistently. Selecting the correct model and the number of clusters, thus providing a novel approach for accurate diagnosis of the disease.

Using this approach, we were able to successfully denoise proteomic spectra and reach up to a 99% total reduction of the number of peaks compared to the original data. In addition, the Bayesian-based approach generated a better classification rate in comparison with other classification algorithms.

This new finding will allow us to apply the Fourier transformation for the selection of the protein profile for each sample, and to develop a novel bioinformatic strategy based on Bayesian clustering for biomarker discovery and optimal diagnosis.

Contacts: bensmail@utk.edu; haoudia@evms.edu

Running head: Bayesian-based clustering of proteomics data
1 INTRODUCTION AND MOTIVATION

1.1 An Approach to proteome analysis using SELDI-Time of Flight-Mass Spectrometry

There is a variety of new methods for proteome analysis. Unique ionization techniques, such as electrospray ionization and matrix-assisted laser-desorption ionization (MALDI), have facilitated the characterization of proteins by Mass Spectrometry (MS) (Karas and Hillenkamp, 1988; Hillenkamp et al., 1991). Surface-enhanced laser desorption-ionization time of flight mass spectrometry (SELDI-TOF-MS), originally described in Hutchens and Yip (1993) overcomes some of the problems associated with sample preparation inherent with MALDI-MS. The underlying principle in SELDI is surface-enhanced affinity capture through the use of specific probe surfaces or chips. Chips with broad binding properties, including immobilized metal affinity capture, and with biochemically characterized surfaces, such as antibodies and receptors, form the core of SELDI analyser. Sample volumes can be scaled down to as low as 0.5 µl, an advantage in cases in which sample volume is limiting. Once captured on the SELDI protein biochip array, proteins are detected through the ionization-desorption, TOF-MS process. A retentate (proteins retained on the chip) map is generated in which the individual proteins are displayed as separate peaks on the basis of their mass and charge (m/z). Wright et al. (1999) demonstrated the utility of the ProteinChip SELDI-MS in identifying known markers of prostate cancer and in discovering potential markers either over- or underexpressed in prostate cancer cells and body fluids. SELDI analyses of cell lysates prepared from pure populations from microdissected surgical tissue specimens revealed differentially expressed proteins in the cancer cell lysate when compared with healthy cell lysates and with benign prostatic hyperplasia (BPH) and prostate intraepithelial neoplasia cell lysates (Cazares et al., 2002). In addition, distinct SELDI protein profiles for each cell and cancer type evaluated, including prostate, lung, ovarian, breast, bladder and head and neck cancer, have been described (Adam et al., 2002; Cazares et al., 2002; Li et al., 2002; Petricoin et al., 2002; Wadsworth et al., 2004; Vlahou et al., 2004). In these studies, protein profiling data is generated by SELDI ProteinChip Array technology followed by analysis utilizing numerous types of software algorithms.

1.2 Clustering and classification methods for large proteomics data sets

Due to the large array of data that is generated from a single proteomic data analysis, it is essential to implement the use of algorithms that can detect expression patterns from such large volumes of data correlating to a given biological/pathological phenotype from multiple samples (Bensmail and Haoudi, 2003). Under normality assumption, covariance matrices play an important role in statistical analysis including clustering. Particularly, covariance matrix provides information on the structure of the data to cluster. The geometrical structure of the dataset is expressed through the eigenvalues and eigenvectors of covariance matrix. Eigenvectors represent the orientation of the data, whereas eigenvalues represent its shape (dispersion). K-means (MacQueen 1967), which uses a minimum distance criterion to find clusters that are not supposed to overlap, and where the number of clusters is given as an input. Self-Organizing maps (Kohonen, 1997), Neural Networks (Bishop, 1995) and other clustering algorithms pay less attention to this structure. K-means put weight on the cluster mean. Kohonen Mapping (SOM) uses a topological structure classification using known weight vectors to initialize the topology of the clusters. Neural Networks, a method borrowed from the computer sciences uses a non linear transformation to cluster data. All the previous methods do not account for the choice of the number of clusters; it has to be specified a priori. Probabilistic models, particularly Bayesian models, offer this possibility, and each also proposes a probabilistic criteria for determining the number of clusters, the geometry and the uncertainty involved in the cluster designation.
1.3 A Bayesian-based clustering method for proteomics data

In order to properly and more efficiently cluster the high dimensional proteomics data obtained through mass spectrometry analysis, and therefore classify different patients groups with higher specificity and sensitivity, we propose a novel clustering approach composed of three components:

(a) Fourier transformation to denoise the data.
(b) Reparametrization of the covariance matrices. We find the geometric characteristic of the data to cluster described by the shape, orientation and volume of each cluster.
(c) Fully Bayesian analysis to account for the probability of classification for each data point and to help include any prior information obtained from the data.

Using this approach, we were able to successfully denoise proteomic spectra and reach up to a 99% total reduction of the number of peaks compared to the original data. In addition, the Bayesian-based approach generated a better classification rate of 96% (misclassification error rate is 4%) in comparison with other classification algorithms, which generated misclassification error rates ranging between 28% and 50%.

2 METHODOLOGY

2.1 Serum samples from HTLV-1-infected patients

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 ATL (Adult T cell leukemia), an aggressive lymphoproliferative disease, from HAM/TSP (HTLV-1-Associated Myelopathy/Tropical Spastic Paraparesis), a chronic progressive neurodegenerative disease. Both diseases are associated with infection T-cells by HTLV-1.

Triplicate serum samples ($n = 70$) from healthy or normal ($n_1 = 38$), ATL ($n_2 = 12$) and HAM ($n_3 = 20$) 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 to 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.

2.2 SELDI mass spectrometry

Serum samples were analyzed by SELDI mass spectrometry as described earlier (Adam et al., 2002). The spectral data generated was used in this study for the development of the novel bayesian clustering approach.

2.3 Data visualization and denoising by a discrete fourier transformation

2.3.1 Visualization

Fourier transformation takes a discrete time series of "$n$" equally spaced values and transforms or converts this series through a mathematical operation into a set of "$n$" complex numbers defined in what is called the frequency domain. If the assumption is made that the time series is made up of oscillating signals of various frequencies plus noise, then in the
frequency domain we can filter out the frequencies of no interest and so minimize the noise content of the data and reduce the dimensionality.

We begin with a data set represented by a $n \times p$ matrix $X$, where $p = 25,196$ is the number of variables (peaks) measured on each sample and $n = 70$ is the number of samples (patients). We first applied the Fourier transform to visualize the data within each class, to denoise it and reduce its dimensionality.

Heuristic-based (ie. pairwise similarity based) clustering algorithms would not have any issues with singularly. For comparison purpose, we used hierarchical clustering algorithm on a distance matrix $(70 \times 70)$. One particular sample was out of the distance range, sample 11. Omitting this observation did not improve the result suggesting that a more robust method is needed to analyze the data.

**Figure 1, here**

A Fourier transform decomposes a sample into two individual sinusoidal components, one real and one imaginary. They are combined together to form the representation of the data in the transformation domain. A Fourier transform takes the form of

$$x(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft} dt$$

where $x(t)$ is the original waveform, $x(f)$ is the Fourier transform of $x(t)$, and $j = \sqrt{-1}$. If $x(t)$ is not periodic, then the Fourier transform is a continuous function of frequency.

In general a function from the transform domain, will be of the form:

$$H(f) = R(f) + jI(f) = |H(f)|e^{j\theta(f)}$$

where $R(f)$ is the real part of the Fourier transform, $I(f)$ is the Imaginary part of the Fourier transform, $|H(f)|$ is the amplitude of the Fourier spectrum ($|H(f)|$ can also be represented as $\sqrt{R(f)^2 + I(f)^2}$), and $\theta(f)$ is the phase angle of the Fourier transform ($\theta(f) = \tan^{-1}[I(f)/R(f)]$).

The fast Fourier transform is a computation algorithm employed to calculate the discrete Fourier transform (DFT). The discrete Fourier transform is a special case of the Fourier transform discussed above, instead of representing the transformed series as an integral, it is represented as a finite summation, as seen below

$$x_{DFT} = \langle x, s_n \rangle = \sum_{k=0}^{N-1} x(k)s_k, \quad k = 0, 1, \ldots, N - 1$$

$$= P_{s_k}x \quad \text{where} \quad \langle s_k \rangle \quad \text{are orthogonal (sinusoidal basis).}$$

We interpret the DFT as the coefficients of projection $P$ of the signal vector $x$ onto the $N$ sinusoidal basis signals $s_k$.

The inverse DFT is then defined as

$$\tilde{x}(n) = \sum_{k=0}^{N-1} P_{s_k}(x), \quad k = 0, 1, \ldots N - 1$$

In summary, the inverse DFT is the reconstruction of the original signal as a superposition of its sinusoidal projections.

### 2.3.2 Denoising and thresholding

Noise and dimension reduction using Fourier transform is based on the fact that each part of a spectra has its own coefficients in the fourier transform. Thus, if we can eliminate the
coefficients belonging to the noise and take the inverse transform of the remaining coefficients, we have denoised data.

Here, we use a thresholding method to specify noise coefficients, i.e. we compare the transformed spectra with a threshold which can be estimated using the soft Donoho’s method (Donoho and Johnstone 1994) given by

\[ y_i = \text{sign}(x_i) \times (x_i - t_i)_+, \text{ where } t_i = C\sigma_i, \quad i = 1, \ldots, n_i \]

where \( x_i \) is the transformed spectra, \( t_i \) is the threshold, \( \sigma_i \) is the standard deviation of the \( i \)th transformed spectra and \( n_i \) is the number of peaks for the \( i \)th sample. We gave \( C \) values of 2 and 3. The thresholding process is equivalent to a projection \( P_{\delta_i}(x) \), where \( P \) is a diagonal matrix \( \text{diag}(p_i) \) where \( p_i = I(|x_i| \geq C\sigma_i), i = 1, \ldots, p \) and \( I(.) \) is the indicator function. The rank of \( P \) is smaller than \( p \) and the reconstructed signal \( y \) obtained by using the inverse discrete Fourier transform, which is provided by R function \( \text{fft}(y, \text{inverse}) \), may still be a very good approximation of the original spectra. When the dimensionality of each sample is very large (25, 196 in this case), thresholding the transformed spectra may need further feature reduction. A P-spline method may be used to further pick important (extremal peaks).

2.4 Choosing the features:

We start by fitting a \( P \) - spline curve \( \hat{y}_i \) to each reconstructed sample \( y_i \) (Hastie and Tibshirani, 1990; Cerioli, Laurini and Corbellini, 2003). For instance, popular cubic spline functions are \( P \)-splines of order 2, penalizing the second derivative of \( \hat{y}_i \). When using a \( P \)-spline of order 4, this leads to an estimate of \( y_i \) with continuous second derivative (Hastie and Tibshirani, 1990).

The R function "predict.smooth.Pspline" provides the first derivative \( \hat{y}_i' \). This information is used to obtain the extremal points of \( y_i \). We perform this step by computing an approximate 95% pointwise confidence interval for the first derivative of \( y_i \) (Silverman 1985). Whenever the lower bound of this interval is greater than zero, we have confidence that \( y_i \) is actually increasing. Similarly, we treat \( y_i \) as decreasing when the upper bound of the interval is negative. Within the interval in which the derivative changes from positive to negative, we have a maximum. Let \( P_i = \{ p_1, \ldots, p_n \} \) be the collection of the number of the selected peaks which are local maxima for each sample. The uniform number of such maxima was defined as \( \text{min} \{ p_1, \ldots, p_n \} \) which is defined to be \( m = 21 \). After transforming, denoising and smoothing, each spectra indicates some metabolites peaks of its source. We use the smoothed peaks remaining as features for the clustering algorithm which will be described in the following section.

2.5 Model-based mixture model

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 (Hartigan, 1975; Gordon, 1999; Kaufman and Rousseeuw, 1990; MacQueen, 1967; Wolfe, 1978; Scott and Symons, 1971; Bock, 1985). Various strategies for simultaneous determination of the number of clusters and the cluster membership have been proposed (Engelman and Hartigan, 1969; Bozdogan, 1993-1994; Bock, 1996; Bensmail and Bozdogan, 2002-2003-2004; Bozdogan, 1999; Bozdogan and Haughton (1998) and Bozdogan and Ueno, 2004).

Mixture models provide a useful statistical frame of reference for cluster analysis. In the theory of finite mixture, the data to be classified are viewed as coming from a mixture of probability distributions, each representing a different cluster, so the likelihood is expressed as

\[ p(\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K | y) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i | \theta_k) \]
where \( \pi_k \) is the probability that an observation belongs to the \( k^{th} \) component or cluster \((\pi_k \geq 0; \sum_{k=1}^{K} \pi_k = 1) \) and \( f_k \) is the density function of each component distribution, \( \theta_k = (\mu_k, \Sigma_k) \) is the underlying parameter involved.

Methods based on this theory performed well in many cases and applications included character recognition (Murtagh and Raftery 1984), tissue segmentation (Banfield and Raftery 1993), minefield and seismic fault detection (Dasgupta and Raftery 1998), application to astronomical data (Bensmail et al. 1997; Roeder and Wasserman 1997; Mukherjee et al. 1998), enzymatic activity in the blood (Richardson and Green 1997); gene expression (Yeung et al. 2001); ovarian cancer detection (Bensmail and Bozdogan 2002-2003).

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; Bensmail and Bozdogan, 2003).

Here, our approach uses a Bayesian mixture model based on a variant of the standard spectral decomposition of \( \Sigma_k \), namely

\[
\Sigma_k = \lambda_k D_k A_k D_k^T
\]

where \( \lambda_k \) is a scalar, \( A_k = \text{diag}(a_{k1}, a_{k2}, \ldots, a_{kp}) \) where \( 1 \geq a_{k1} \geq \ldots a_{kp} > 0 \), and \( D_k \) is an orthogonal matrix for each \( k = 1, \ldots, K \).

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 \( y_i \), \((i = 1, \ldots, n; y_i \in \mathbb{R}^p)\) to be classified arise from a random vector \( X \) with density \( p(0, \pi|Y = y) \) as in (1), where \( f_k(\cdot|\theta_k = (\mu_k, \Sigma_k)) \) is the multivariate normal density function, \( \mu_k \) is the mean and \( \Sigma_k \) is the covariance matrix for the \( k^{th} \) group. Vector \( \pi = (\pi_1, \ldots, \pi_K) \) is the mixing proportion \((\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1) \). We are concerned with Bayesian inference about the model parameters \( \theta \), \( \pi \) and the classification indicators \( v \). Markov Chain Monte Carlo methods (MCMC) provide an efficient and general recipe for Bayesian analysis of mixtures. Given a classification vector \( v = (v_1, \ldots, v_n) \), we use the notation \( n_k = \#\{i : v_i = k\} \) for the number of observations in cluster \( k \), \( \bar{y}_k = \sum_{i;v_i=k} y_i/n_k \) for the sample mean vector of all observations in the cluster \( k \), and \( W_k = \sum_{i;v_i=k} (y_i - \bar{y}_k)(y_i - \bar{y}_k)^t \) for the sample covariance matrix.

We use conjugate priors for the parameters \( \pi \) and \( \theta \) \((\lambda_k, D_k, A_k, \pi)\) of the mixture model. The prior distribution of the mixing proportions is a Dirichlet distribution

\[
\pi_1, \ldots, \pi_K \sim \text{Dirichlet}(\beta_1, \ldots, \beta_K),
\]

with joint distribution

\[
p(\pi) = \frac{\Gamma(\beta_1 + \ldots + \beta_K)}{\Gamma(\beta_1) \ldots \Gamma(\beta_K)} \pi_1^{\beta_1-1} \ldots \pi_K^{\beta_K-1}
\]

The prior distributions of the means \( \mu_k \) of the mixture components conditionally on the covariance matrices \( \Sigma_k \) are Gaussian

\[
\mu_k | \Sigma_k \sim \mathcal{N}(\bar{y}_k, \Sigma_k/\tau_k)
\]

with known scale factors \( \tau_1, \ldots, \tau_K > 0 \) and locations \( \xi_1, \ldots, \xi_K \in \mathbb{R}^p \). The conjugate prior distribution of the covariance matrices depends on the model, and will be given for each model in turn as detailed in Table 1.

Table 1 here
probabilities \( p_{ik} = P(v_i^{(o)} = k|\pi, \theta), \ k = 1, \ldots, K \) conditional on the current values for \( \pi^{(o)} \) and \( \theta^{(o)} \),

\[
p_{ik} = \pi_k f_k(y_i | \mu_k^{(o)}, \Sigma_k^{(o)}) / \sum_{k=1}^{K} \pi_k f_k(y_i | \mu_k^{(o)}, \Sigma_k^{(o)}) \quad (i = 1, \ldots, n).
\]

There might be classes which are empty. To solve this problem, we affect the closest observation to the empty class.

2. Simulate the vector \( \pi^{(t+1)} = (\pi_1^{(t+1)}, \ldots, \pi_K^{(t+1)}) \) of mixing proportions from its posterior distribution, namely

\[
\pi^{(t+1)} \sim \text{Dirichlet}(\beta_1 + \sum_{i=1}^{n} \# \{v_i^{(t+1)} = 1\}, \ldots, \beta_K + \sum_{i=1}^{n} \# \{v_i^{(t+1)} = K\})
\]

With \( \beta_k \) the known prior parameters of the Dirichlet distribution.

3. Simulate the parameter \( \theta^{(t+1)} \) of the model from the posterior distribution \( \theta | v^{(t+1)} \).

4. Iterate the steps 1 to 3.

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 \( \theta \), was shown 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 values of \( \theta \) and \( \pi \), and important properties as the central limit theorem or the law of the iterated logarithm are then satisfied. In the following, we describe and estimate parameters of the four proposed model we used in this paper:

(a) Model [\( \Sigma \)] (similar ellipsoidal clustering)
Using this model, we suppose that all clusters have the same volume, same shape and same orientation (\( \Sigma = \lambda DAD^\top \)). The prior distribution of the parameter \( \mu_k \) is given in (2). The prior distribution of \( \Sigma \) is an inverse wishart distribution with degrees of freedom \( m_\Sigma \) and sample variance \( \Psi_0 \). Then the posterior distribution of \( (\mu_k, \Sigma, v) \) is a multivariate normal distribution with mean \( \bar{\xi}_k = (n_k \bar{y}_k + \tau_k \xi_k) / (n_k + \tau_k) \) and covariance matrix \( \Sigma / (n_k + \tau_k) \) and the posterior distribution of \( \Sigma | \mu, y \), has the following inverse Wishart distribution

\[
W_{p-1}^{-1} \left( n_k + m_\Sigma, \Psi_0 + W + \sum_k n_k \lambda (\bar{y}_k - \xi_k)(\bar{y}_k - \xi_k)^\top / (h_k + \tau_k) \right)
\]

(b) Model [\( \Sigma_k \)] (different ellipsoidal clustering)
In this case, we suppose that clusters have different shape, different volume and different orientation (\( \Sigma = \lambda_k D_k A_k D_k^\top \)). If the prior distribution of \( \Sigma_k \) is an inverse wishart distribution with degrees of freedom \( m_k \) and a sample variance \( \Psi_k \), then the posterior distribution of \( \Sigma_k | \mu, x \), has the following inverse Wishart distribution

\[
W_{p-1}^{-1} \left( n_k + m_k, \Psi_k + W_k + n_k \lambda (\bar{y}_k - \xi_k)(\bar{y}_k - \xi_k)^\top / (n_k + \tau_k) \right)
\]

(c) Model [\( \lambda \)] (similar spherical clustering)
This model assume that the clusters are spherical with the same volume (\( \lambda \)). As shown above, the posterior distribution of \( (\mu_k, \lambda, v) \) is a multivariate normal distribution with mean \( \bar{\xi}_k = (n_k \bar{y}_k + \tau_k \xi_k) / (n_k + \tau_k) \) and covariance matrix \( \lambda (n_k + \tau_k) I \). As \( \lambda \) is a scale measure, we use an inverse Gamma distribution \( G^{-1}_a(m_\lambda/2, \rho_\lambda/2) \) as a prior with scale and shape parameters \( m_\lambda \) and \( \rho_\lambda \). The posterior distribution of \( \lambda | \mu, y \) is an inverse Gamma distribution.
\[ G_a^{-1} \left( \frac{1}{2} (n_k + pm_k), \left\{ p_k + tr(W_k + \sum_k \frac{n_k \xi_k}{n_k + \xi_k} (\tilde{y}_k - \xi_k) (\tilde{y}_k - \xi_k)' \right\} / 2 \right) \]

(d) Model [\( \lambda_k \)] (different spherical clustering)
This model assume that the clusters are spherical with different volume (\( \lambda_k \)). If the prior distribution of \( \lambda_k \) is an inverse Gamma \( \sim G_a^{-1}(m_k/2, \rho_k/2) \) with scale and shape parameters \( m_k \) and \( \rho_k \) for each cluster, then the posterior distribution of \( \lambda_k | x \) is an inverse Gamma distribution
\[ G_a^{-1} \left( \frac{1}{2} (n_k + pm_k), \left\{ p_k + tr(W_k + \sum_k \frac{n_k \xi_k}{n_k + \xi_k} (\tilde{y}_k - \xi_k) (\tilde{y}_k - \xi_k)' \right\} / 2 \right) \]

After describing different models involved in the clustering algorithm, we propose a bayesian approach for selecting the number of clusters. Do we have two clusters, three, four, etc? We also need to specify which one of the models described above offer a good number of clusters. Models described above uses the geometrical specification of the clusters given the type of data to be analyzed. To select the number of clusters and the model which fit well the data, we use a bayesian score function based on the integrated likelihood which will be described in the following section.

2.6 Model selection
After describing the models of interest and the analytical forms of the posterior densities, approximate Bayes factors and information based criterias will be used in this section to compare models, in the strategy of identifying the linear, quadratic or spherical models and to identify the number of clusters.

For a deterministic approach, popular criteria’s were proposed and used. This includes Akaike Information Criteria (AIC) (Akaike, 1973) defined as
\[ AIC(M_k) = -2 \log L(\hat{\theta}_k, M_k) + d \]
Bayesian Information Criterion (BIC) introduced by Schwarz (Shwartz, 1978) defined as
\[ BIC(M_k) = -2 \log L(\hat{\theta}_k, M_k) + d \log(n_k) \]
and Information Complexity Criterion (ICOMP) (Bozdogan 1994) defined by
\[ ICOMP(M_k)_{IFIM} = -2 \log L(\hat{\theta}, M_k) + s \log \left( \frac{\bar{\lambda}_g}{\bar{\lambda}_o} \right) \]
where
\[ s = \text{rank}([\text{cov}(\hat{\theta})]), \quad \bar{\lambda}_o \] is the arithmetic mean and \( \bar{\lambda}_g \) is the geometric mean of the eigenvalue.
Models having a smaller Information criteria should be favored as this indicates a better fit and a lower degree of model complexity. Note that ICOMP now looks in appearance like Rissanen’s (1978) MDL, and Schwarz’s (1978) Bayesian criterion BIC, except for using \( \log \left( \frac{\bar{\lambda}_o}{\bar{\lambda}_g} \right) \) instead of using \( \log(n) \), where \( \log(n) \) denotes the natural logarithm of the sample size \( n \). A model with minimum ICOMP is chosen to be the best among all possible competing alternative models. The greatest simplicity, that is zero complexity, is achieved when \( \text{Cov}(\hat{\theta}) \) is proportional to the identity matrix, implying that the parameters are orthogonal and can be estimated with equal precision. In this sense, parameter orthogonality, several forms of parameter redundancy, and parameter stability are all taken into account.

In the stochastic approach, the Bayes factor provides a measure of whether the data have increased or decreased the odds of a model to choose. To choose a model \( M_2 \) against \( M_1 \), the approximate Bayes factor is given by
\[ BF_{1,2} = p(y|M_2)/p(y|M_1) \]
where \( p(y|M_k) = \int p(y|\theta_k)p(\theta_k|M_k)\text{d}\theta_k \), \( \theta_k \) is the vector of parameters of \( M_k \) and \( p(\theta_k|M_k) \) is its true prior density. Often it is difficult to compute the integral involved in \( p(y|M_k) \) and therefore some approximation methods are used. One possibility is to approximate the integral by Laplace’s method using the normal approximation. This is simple to calculate and proven to give accurate estimates (Gelfand and Dey, 1994; Lewis and Raftery, 1994; Bensmail et al.,
In this case, Bayes factor is defined as

\[ BF_{1,2} = \frac{p(y|M_2)}{p(y|M_1)} = \frac{|-H_2^{-1}(\tilde{\theta}^{(2)})|^{1/2}p(y|\tilde{\theta}^{(2)})p(\tilde{\theta}^{(2)})(2\pi)^{p_2}}{|-H_1^{-1}(\tilde{\theta}^{(1)})|^{1/2}p(y|\tilde{\theta}^{(1)})p(\tilde{\theta}^{(1)})^{p_1}}, \]

where \( \tilde{\theta}^{(i)} \) is the posterior mode of \( \theta^{(i)} \) (denoting the parameters \( \mu \) and \( \Sigma \) of the model \( M_i \)), and \( H^{(i)} \) is the Hessian of \( h(\theta) = \log p(y|\theta)p(\theta) \), evaluated at \( \theta = \tilde{\theta}^{(i)} \). The terms \( \tilde{\theta} \), and \( | -H^{(i)}(\tilde{\theta})|^{-1} \) are calculated using the Gibbs sampler output. The likelihood at the approximate posterior mode is

\[ p(y|\tilde{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tilde{f}_{k} k(y_i|\tilde{\mu}_k, \tilde{\Sigma}_k) \]

which is then substituted into equation (3) to obtain 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 \( 1, \ldots, K \) for all models. By convention, \( \log(BF_{12}) < 2 \) represents weak evidence, differences between 2 and 6 represent positive evidence, differences between 6 to 10 represent strong evidence, and differences > 10 represent very strong evidence (Jeffrey, 1961).

### 3 RESULTS

#### 3.1 SELDI mass spectrometry data set

Available for cluster analysis were the spectrums of mass and intensity from 70 different patients. healthy or normal (\( n_1 = 38 \)), ATL (\( n_2 = 12 \)) and HAM (\( n_3 = 20 \)) patients were processed as shown in figure 2. Representative duplicate and rescaled spectra from different groups (Normal, ATL and HAM) are shown in figure 3. The spectra presented in duplicate showed the high reproducibility of our proteomic profiling approach using SELDI-TOF-MS. This rescaled figure shows clearly the peaks as features which will be used for clustering. There is a random distribution of the peaks. The largest peak of the normal sample might not be the case for the Ham or ATL.

**Figure 2, here**

**Figure 3, here**

#### 3.2 Data visualization by a Discrete Fourier transform

To better visualize the proteomics data, we used discrete fourier transform. We applied multivariate discrete fourier transformation to transform intensity spectra alone and the combined mass and intensity, as shown in figures 4 and 5, respectively. These images show clear distinction between the three different group: Normal, ATL and HAM which the medical researcher expects to find using a powerful and accurate clustering method.

**Figure 4, here**

Clearly, the image in the left panel summarizing the normal group, has a ellipsoidal core in the middle with two circular extrems on the top and the bottom surrounding the middle ellipse. The second image in the middle panel summarizing the ATL group is smoother, no circular or ellipsoidal form is apparent. The third image on the right panel summarizing HAM group has the same middle form as for the normal group but is smoother on the extrems (figure 4). Generally, those similarities hold for most of the 70 samples except for few which may be interesting to look at later.

**Figure 5, here**

For each patient there were two spectra available for analysis, the mass and the intensity.
of the sample. While the length of these two spectra are equal to \( p = 25,196 \) for each sample. Variability and noises are highly expressed in the spectra data. Therefore any clustering algorithm will fail to detect clusters if the data are not denoised and processed first.

Using discrete fourier transform we visualized the data so that the distinction among the three groups of patients, NOR, ATL and HAM become apparent and each group has its own specific features.

### 3.3 Data denoising by a discrete fourier transform

Threshold with Fourier transform was used to remove the irrelevant peaks and then reduce the length of the spectra of different patients (Figure 6). A spectra from the normal group is displayed in figure 6A. We first applied a nonlinear smoothing form of the original data using the running medians method "smooth" function in SPLUS (Figure 6B). We noticed that using a traditional smoothing method, we failed to detect noises and then reduce the number of peaks. Therefore, we decided to use the thresholding fourier transform as shown in figure 6C. Frequencies, which are the projection of the original data into the orthogonal sinusoidal basis, are displayed. We then reconstruct the data after thresholding and reducing the frequencies using the Donoho’s approach described earlier (Figure 6D). The length of the original data was reduced from \( p = 25,196 \) to a minimum of 84 to and maximum of 120 peaks, which is a 99% total reduction of the number of peaks from the original data.

Figure 6, here

It is readily apparent from figure 6 that there are distinct differences in the nature of the peaks from the intensity spectra. The difficulty lies in identifying their position. It does not suffice to say that the fiftieth observation will represent a peak because the nature of the fiftieth observation is very different depending on the length of the spectra being examined. Therefore we seek a methodology that will allow us to either identify these areas of importance by partitioning the spectrum, or achieve classification ability utilizing all information currently available.

These disparities make traditionally used methodologies difficult to implement. One cannot merely say that there are \( x \) –peaks or \( x \) –variables in the study and implement a clustering algorithm to classify the data using the peaks as feature. The graphical representation of the mass and intensity spectra are vastly different.

At the last phase, we identify areas of importance by shrinking all the series down to uniform length using the smoothing P-spline to obtain extremal points, which reduces the length down to 21 peaks and then apply the proposed Bayesian algorithms to find the clusters. Figure 7 illustrates the intensity spectra for one patient from each category, and figure 8 represents the mass spectra.

Figure 7, here

Figure 8, here

Using discrete fourier transform, we successfully denoised spectral data up to 99% total reduction of the number of peaks from the original data.

Now that the data are summarized in a non singular matrix of dimension \((70 \times 21)\). Any clustering algorithm may be applied to classify the transformed data.

### 3.4 Comparative clustering analysis and Bayesian model development

We applied the bayesian clustering algorithm and succeeded in detecting the clusters and their geometrical representation and also provided the uncertainty of the classification for each sample. We compared the performance of our proposed bayesian clustering to some traditional classification and clustering algorithms including K-means, SOM (Kohonen...
mapping), and LDA (Linear Discriminant Analysis). We found that the Kohonen Self-Organizing Map is not an effective technique for clustering. The K-means clustering algorithm leads one to believe that there are only two clusters present in the data. Given the scientific information that indicates that both ATL and HAM patients are infected from the same virus, these results are not completely surprising. The linear discriminant analysis do not predict a clear three supervised clustering.

Using the mass only, 500 iterations of the Gibbs sampler was found necessary for stability. Convergence was immediate. Similar results were obtained for intensity and for intensity and mass. The proposed model \([\lambda_k I]\) (model 2) and the correct number of groups (2 clusters), are strongly favored using Bayes factor. Bayes factor BF scored best for the model \([\lambda_k I]\) with two components (Table 3), which means that data proposes two spherical groups with different volume (ICOMP remains consistent with Bayes factor in choosing the same number of clusters but it uses the model 1 \([\lambda I]\), Table 2). The posterior modes of the cluster volume for the preferred model is \(\hat{\lambda}_1 = 2.009, \hat{\lambda}_2 = 1.91\). The error rate of misclassification using mass only obtained by the BF-second optimal model is equal to 9\% (Table 5).

Using the intensity only, the proposed model \([\Sigma_k]\) (model 4) and the correct number of groups (3 clusters), are strongly favored. Bayes factor BF scored best for the model \([\Sigma_k]\) with three components (Table 3), which means that data proposes three different ellipsoidal classes with the different volume and different direction to both axes (ICOMP scored for two clusters using the model 4 \([\Sigma_k]\), Table 2). The error rate of misclassification using intensity alone obtained by the BF-optimal model is equal to 7\% (Table 5).

Using both intensity and mass, the proposed model \([\lambda_k I]\) (model 2) and the correct number of groups (3 clusters), are strongly favored. Bayes factor BF scored best for the model \([\lambda_k I]\) with three components (Table 3), which means that data proposes three different spherical classes with different volume (ICOMP was consistent with Bayes factor in choosing the number of clusters and the same model 2 \([\lambda_k I]\), Table 2). The posterior modes of the clusters volume are; \(\hat{\lambda}_1 = 1.064, \hat{\lambda}_2 = 1.0655, \hat{\lambda}_3 = 1.06418\) which are close to the true values (Figure 9 and 10).

In figure 11 we show for each projected data point (projection in the first and second principal components basis) its class assignment resulting from the maximum a posteriori rule with three groups. The error rate of misclassification using both mass and intensity obtained by the BF-optimal model is equal to 4\% (Table 5).

Perhaps one of the greatest advantages of the bayesian approach is that it fully assesses uncertainty about group membership, rather than merely giving a single "best" partition. The uncertainty is measured by \(U_i = \min_{k=1,\ldots,K} \left(1 - \Pr[v_i = k|\text{Data}]\right)\).

In the three cases, SOM, K-means, and LDA performed poorly in comparison with the Bayesian model. The Bayesian-based approach generated better classification rate of 96\% (misclassification error rate is 4\%) in comparison with other classification algorithms generating a misclassification error rate reaching 28\% to 50\% in the best case.

3.5 Simulated data:
To show the performance of the proposed algorithm, we simulated 150 observations (three
groups of size 50). Observations within each group are simulated from a uniform distribution in \([0, 60]\). 1000 observations were simulated for the first group, 1000 for the second group and 1000 for the third group. We added white noises \(\varepsilon \sim \mathcal{N}(0, \sigma_k^2 = \frac{1}{k}, k = 1, 2, 3)\), 6200 for the first group, 420 for the second group and 6460 for the third group (Figure 12). We based this scheme on the working paper by Garrett and Parmigiani (2003), Fraley and Raftery (1998) and Lee et al. (2000).

In practice, the normal distribution have proven successful in capturing the nature of genomics and proteomics expression. There are many distributions which would likely achieve the same goals. Our reason for choosing the above distributions are partly mathematical convenience and partly due to the nature of the genetic and proteomics data. For example, it can be assumed in many cases that the error associated with measuring proteomics expressions follows a Gaussian distribution, justifying our use of the normal distribution for noises. In our applied setting, the uniform distribution naturally lends itself to the case of proteomics expressions. In cancer applications, differential are thought to be caused by the failure of biological mechanisms. As a result, the observed expression may take a broad range of values.

After denoising and thresholding the series, the length of the simulated data was reduced to a maximum of 173 peaks, which is a 97% total reduction of the number of peaks. Shrinking all the series down to a uniform length using the smoothing P-spline, we reduce the length down to 17 peaks. The Bayesian algorithm was applied to find the clusters and the best model which describes the data. Bayes factor \(BF\) scored best for the model \([\lambda_k l]\) with three components \(BF = 1024\), while ICOMP scored best for the model \(\Sigma_k = [\lambda_k D_k A_k D_k^t]\) with three clusters. This means that both criterion agree on proposing groups with different covariance matrices. Bayes factor chooses a parsimonious model (spherical) while ICOMP proposes a more complex model (general model).

4 DISCUSSION

Proteomic technologies hold great promise toward identification of proteins to serve as biomarkers for cancer diagnosis, to monitor disease progression, and to serve as therapeutic targets.

Protein expression profiling is used to identify a panel of proteins expression events that are associated with the disease. The high dimensionality of the data obtained through SELDI mass spectrometry underscores the need for developing an algorithm capable of analyzing such high volume data to develop an efficient and reproducible classifier. In order to achieve this goal, we have implemented the use of Bayesian-based Fast fourier algorithm to denoise the data and asked whether the processed (denoised) data will be more efficiently clustered.

With our data set, the fourier transformation bayesian-based approach not only showed an extremely good performance, but proposed the right number of clusters and their geometrical configuration using Bayes factor with Laplace approximation. The diagonal model proposed by Bayes factor performed well on the mass spectral data. This means that the mass of the proteome is summarized by circular clusters. The number of clusters proposed based on the mass only, is two, which might be surprising but not contradictory. This indicates that the mass profiles of the proteome of patients who are infected by the same virus, HTLV-1, are similar. When the approach was used on the intensity data set, Bayes factor generated different ellipsoidal models, for different covariance matrices, and proposed three clusters. The three groups have a superior similarity to the real data sets (Normal, ATL and HAM) with an error rate of 7%. This means that the intensity profile of the proteome can distinguish between the normal patients, the diseased patients and within the two diseased patients, it
recognizes the ATL and HAM and separate them in two different clusters. When both data are combined, this gives a clear superior performance to our clustering approach. The model proposed is a diagonal covariance (parsimonious) and the number of clusters is three and an error rate of misclassification is down to 4%.

In the majority of the clustering cases we have analyzed, we found that the self organizing map converges to only one cluster. Our results also indicate that using the classical methods of classification, the medical researchers may be misled by the outcome of two clusters present in this data set, but as many medical researchers suspect, there may in fact be three clusters present in this data set as the bayesian approach specified. The Bayesian-Fourier model-based approach has superior performance, consistently, selecting the correct model and the number of clusters, thus providing a novel approach for accurate diagnosis of the disease.

The high specificity obtained using this approach represents a significant advancement in the clustering of high dimensional data especially when more than two patient groups are considered. Classifying proteomics data generated from studies containing more than two patient groups represent a serious challenge and the outcome of such classification is usually a low specifies and sensitivity rate. To boost the capabilities of this algorithm and enhance the rate of correct classification, one may instead use a non parametric approach to cluster the Fourier transformed proteomic data. Use of the normality assumption was strongly emphasized and pushed on the transformed data, which causes a satisfactory but still not perfect score especially when the mass is used instead of the intensity.

5 CONCLUSION

The development of efficient and powerful bioinformatics tools for accurate and rapid analysis of large proteomic data sets is very beneficial in order to accurately and rapidly interpret proteomics data generated from highthrouput protein expression profiling applications.

We showed that fourier transformed data can play a major role in eliminating noise from proteomics data. It can also be a major tool for reducing the dimensionality of the proteomics data. Even though the proteomics data are not necessarily supposed to be Gaussian, the use of fourier transform enhance it and the bayesian based-model approach produces a good results by comparison to the heuristic methods. The diagonal model and ellipsoidal model produces high quality classification in all the cases: mass, intensity and mass and intensity. The fact that the diagonal model proposed two clusters for the mass data, does not mean that the model is poor. It certainly means that the use of the mass data set as a basis for clustering diseased patients is not reliable. The two boundary of the clusters is not clear on the mass.

Given the distribution of the data collected from patient samples, the most appropriate computational tool for clustering and classification would be based on an unsupervised method. Therefore, we have developed a Bayesian-based classification/clustering algorithm which is to date the most powerful method offering unsupervised classification. Our results suggests the importance of the use of the fourier transform to process the proteomics data and the use of the bayesian model-based approach to find the cluster. This will enable the identification of state- and stage-specific proteins, and therefore may lead to the identification of disease biomarkers.

Developing such enhanced classification algorithms will help to strengthen the interdependence between proteomic technologies and bioinformatics tools to better manage, classify and interpret large proteomics data sets to better understand the complexities between the normal and abnormal cell proteome. We envision that the methodology we have developed and evaluated will be applicable to the study of the proteomes of various biological
systems including environmental hazards, infectious agents (bioterrorism), and cancers.

An improvement of the approach can be developed in many direction. The most urgent and important one is how to find the boundary between the diseased clusters using the mass of the proteome only. This needs a gaussian-like mixture model. For example, a smoothing Gaussian kernel-based mixture model can be an alternative to the gaussian-based mixture assumption. This may be a smoothing function for finding the boundary between the two diseased group of patients (HAM and ATL).

We have previously proposed a non-parametric Bayesian clustering approach to cluster nonlinear data generated from financial institution (Bensmail and Bozdogan 2003). The method uses mixture of multivariate kernel distribution. The data to be classified is supposed to be drawn from a mixture kernel. We used particularly multivariate Gaussian kernel which involves a covariance matrix and a window width to smooth the density function of the observation. In that study, geometric characteristics of the clusters is not valid but the geometric characteristics of the neighbor of each data to classify is considered. The study performs well on real and simulated data. We intend to apply the same methodology to large scale proteomics data to successfully generate meaningful clusters.

ACKNOWLEDGEMENTS

We would like to thank Robert Mee and Lisa Cazares for critical reading of the manuscript. This work was supported by the SRGP Award by the College of Business, University of Tennessee in Knoxville to the first author, and by the Leukemia Lymphoma Society and the National Institutes of Health to the last author.

REFERENCES


Biometrics, 27, 387–397.
LEGEND TO FIGURES

Figure 1: Hierarchical clustering of the SELDI intensity spectra from normal, HAM and ATL patients. Left panel shows no pattern and an unusual observation no 11. Right panel shows no pattern even when the unusual observation no 11 is omitted.

Figure 2: Unprocessed representative SELDI-MS spectral data. SELDI-MS spectra for Normal (A), ATL (B) and HAM patients (C) with a high dimensionality and a lot of noise.

Figure 3. SELDI-MS mass/intensity spectra derived from Normal (A), ATL (B) and HAM (C) patients serum samples. Molecular mass (Da) is presented on the x-axis. Intensity is presented on the y-axis. A representative profile of molecular ions ranging from 6000 to 10000 Da is shown. A duplicate of each spectra was displayed to show the high reproducibility of our results.

Figure 4. Intensity Fast Fourier Transforms
Transformed series for the intensity alone for Normal (A), ATL (B) and HAM (C) patients samples using multivariate discrete fourier transformation. The x-axis and the y-axis summarize the frequencies of the intensity transform.

Figure 5. Intensity and Mass Fast Fourier Transforms
Transformed series of the combined mass and intensity for Normal (A), ATL (B) and HAM (B) patients samples using multivariate discrete fourier transformation. The x-axis and y-axis summarize the frequencies of the mass transform.

Figure 6: Original Intensity, fourier transform, and reconstructed intensity of SELDI spectra from a normal patient
(A) A spectra from the normal group. (B) A nonlinear smoothing form of the original data using the running medians method. (C) the thresholding fourier transform. (D) Reconstructed data after thresholding and reducing the frequencies using the Donoho’s approach.

Figure 7: Reduced intensity spectra using Bayesian algorithms.
The intensity spectra were reduced for one normal (A), ATL (B) and HAM (C) patient.

Figure 8: Reduced mass spectra using Bayesian algorithms.
The intensity spectra were reduced for one normal (A), ATL (B) and HAM (C) patient.

Figure 9: Gibbs sampler output for 500 iterations for the first component mean of different clusters.
(A) Normal, (B) ATL, and (C) HAM patient

Figure 10: Gibbs sampler output for 500 iterations for the volume of the three clusters
(A) Normal, (B) ATL, and (C) HAM patient

Figure 11: Projection of the intensity spectra and its classification.

Table 1: Characteristics of the four models.
The models are: (Σ ) Linear, ( Σ k ) quadratic, (λ I ) spherical with same volume and (λ k I ) spherical with different volume.

Table 2: ICOMP scores for intensity, mass and both.
ICOMP was scored for different number of clusters and different models.

Table 3: Bayes factor scores for intensity, mass and both.
Bayes factor was scored for different number of clusters and different models

Table 4: Percentage of well classified data for mass and intensity and both for three different groups of patients samples.

Table 5: Cross-classification table giving the clustering results for the intensity spectra data.
FIGURES AND TABLES

Dendogram for all patients

Dendogram for patients without no 11

hclust ("average")

hclust ("average")

Figure 1: Hierarchical clustering of HTLV-1 data
Figure 2: Three samples, one normal and two diseased (ATL and HAM)
Figure 3: A dupliacte of each spectra: Normal, HAM and ATL
Figure 4: Intensity Fourier transform
Figure 5: Intensity and mass Fourier transform
Figure 6: Original, transformed and reconstructed data.
Figure 7: Reduced intensity spectra for Normal, ATL and HAM
Figure 8: Reduced mass spectra for Normal, ATL and HAM
Figure 9: Time series of the first 500 Gibbs sampler iterations for the clusters mean.
Figure 10: Time series of the first 500 Gibbs sampler iterations for the clusters volume.
Figure 11: Projection of the intensity spectra and its classification.
Figure 12: Three samples from simulated data
### TABLES:

<table>
<thead>
<tr>
<th>Models</th>
<th>Volume</th>
<th>Shape</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda I )</td>
<td>same</td>
<td>same(spherical)</td>
<td>undefined</td>
</tr>
<tr>
<td>( \lambda_k I )</td>
<td>different</td>
<td>same(spherical)</td>
<td>undefined</td>
</tr>
<tr>
<td>( \lambda DAD^t )</td>
<td>same</td>
<td>same</td>
<td>same</td>
</tr>
<tr>
<td>( \lambda_k D_k A_k D_k^t )</td>
<td>different</td>
<td>different</td>
<td>different</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of the four models

<table>
<thead>
<tr>
<th>Clusters/Models</th>
<th>Mass</th>
<th>Intensity</th>
<th>Mass and Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3351 [( \Sigma )]</td>
<td>3316 [( \Sigma )]</td>
<td>3292 [( \lambda I )]</td>
</tr>
<tr>
<td>2</td>
<td>3242 [( \lambda I )]</td>
<td>3122 [( \Sigma_k )]</td>
<td>3238 [( \Sigma )]</td>
</tr>
<tr>
<td>3</td>
<td>3273 [( \Sigma )]</td>
<td>3216 [( \Sigma )]</td>
<td>3231 [( \lambda_k I )]</td>
</tr>
<tr>
<td>4</td>
<td>3262 [( \lambda_k I )]</td>
<td>3269 [( \lambda I )]</td>
<td>3246 [( \Sigma_k )]</td>
</tr>
<tr>
<td>5</td>
<td>3356 [( \lambda I )]</td>
<td>3272 [( \lambda_k I )]</td>
<td>3268 [( \Sigma )]</td>
</tr>
<tr>
<td>6</td>
<td>3392 [( \Sigma_k )]</td>
<td>3296 [( \lambda_k I )]</td>
<td>3372 [( \Sigma_k )]</td>
</tr>
</tbody>
</table>

Table 2: ICOMP scores for different clusters and different models

<table>
<thead>
<tr>
<th>Clusters/Models</th>
<th>Mass</th>
<th>Intensity</th>
<th>Mass and Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>240 [( \Sigma )]</td>
<td>236 [( \Sigma )]</td>
<td>239 [( \lambda I )]</td>
</tr>
<tr>
<td>2</td>
<td>204 [( \lambda_k I )]</td>
<td>204 [( \Sigma_k )]</td>
<td>213 [( \Sigma )]</td>
</tr>
<tr>
<td>3</td>
<td>210 [( \Sigma )]</td>
<td>216 [( \Sigma )]</td>
<td>211 [( \lambda_k I )]</td>
</tr>
<tr>
<td>4</td>
<td>231 [( \lambda_k I )]</td>
<td>254 [( \lambda I )]</td>
<td>226 [( \Sigma_k )]</td>
</tr>
<tr>
<td>5</td>
<td>245 [( \lambda I )]</td>
<td>253 [( \lambda_k I )]</td>
<td>229 [( \Sigma )]</td>
</tr>
<tr>
<td>6</td>
<td>249 [( \Sigma_k )]</td>
<td>256 [( \lambda_k I )]</td>
<td>231 [( \Sigma_k )]</td>
</tr>
</tbody>
</table>

Table 3: Bayes factor scores for different clusters and different models
<table>
<thead>
<tr>
<th>Clinical</th>
<th>Predicted</th>
<th>Normal</th>
<th>HAM</th>
<th>ATL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>37</td>
<td>1</td>
<td>0</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>HAM</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>ATL</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>% correct</td>
<td><strong>0.97</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.96</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Cross-classification table giving the clustering results for the mass and intensity spectra data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mass</th>
<th>Intensity</th>
<th>Mass&amp;Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>66%</td>
<td>69%</td>
<td>72%</td>
</tr>
<tr>
<td>SOM</td>
<td>50%</td>
<td>54%</td>
<td>56%</td>
</tr>
<tr>
<td>K-Means</td>
<td>59%</td>
<td>53%</td>
<td>61%</td>
</tr>
<tr>
<td>Model-Based</td>
<td><strong>91%</strong></td>
<td><strong>93%</strong></td>
<td><strong>96%</strong></td>
</tr>
</tbody>
</table>

Table 5: Percentage of good classification for mass, intensity and mass & Intensity samples.