University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working
Paper Series

Year 2004

Paper 46

Finding cancer subtypes in microarray data using random projections

Debashis Ghosh*

*University of Michigan, debashis.ghosh@ucdenver.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/umichbiostat/paper46>

Copyright ©2004 by the author.

Finding cancer subtypes in microarray data using random projections

Debashis Ghosh

Abstract

One of the benefits of profiling of cancer samples using microarrays is the generation of molecular fingerprints that will define subtypes of disease. Such subgroups have typically been found in microarray data using hierarchical clustering. A major problem in interpretation of the output is determining the number of clusters. We approach the problem of determining disease subtypes using mixture models. A novel estimation procedure of the parameters in the mixture model is developed based on a combination of random projections and the expectation-maximization algorithm. Because the approach is probabilistic, our approach provides a measure for the number of true clusters in a given dataset. We illustrate our approach with applications to both simulated and real microarray data.

Finding cancer subtypes in microarray data using random projections

Debashis Ghosh

Department of Biostatistics, University of Michigan

Corresponding author:

Debashis Ghosh, Ph.D.

Department of Biostatistics

School of Public Health, University of Michigan

1420 Washington Heights, Room M4057

Ann Arbor, Michigan 48109-2029

Phone: (734) 615-9824

Fax: (734) 763-2215

Email: ghoshd@umich.edu

Abstract

One of the benefits of profiling of cancer samples using microarrays is the generation of molecular fingerprints that will define subtypes of disease. Such subgroups have typically been found in microarray data using hierarchical clustering. A major problem in the interpretation of the output is determining the number of clusters. We approach the problem of determining disease subtypes using mixture models. A novel estimation procedure of the parameters in the mixture model is developed based on a combination of random projections and the expectation-maximization algorithm. Because the approach is probabilistic, our approach provides a measure for the number of true clusters in a given dataset. We illustrate our approach with applications to both simulated and real microarray data.

Keywords: Expectation-Maximization algorithm; gene expression data; High-dimensional data; Mixture models.

Introduction

The development of oligonucleotide and cDNA microarrays has permitted the monitoring of transcript mRNA expression on a global basis. A variety of statistical issues in the analysis of microarray data have been recently addressed, ranging from low-level issues such as preprocessing and normalization of the data [25,27] to differential expression of genes in experimental conditions [13,19].

The problem addressed in this paper is class discovery in cancer settings. Researchers have shown that using microarray technology, one can define disease subtypes in cancers such as cutaneous melanoma [7] and B-cell lymphoma [1]. The main technique for class discovery has been hierarchical clustering [14]. Determining which clusters are “real” and which are not poses a major problem in the consideration of clustering results. Several methods exist for a related problem, that of estimating the number of clusters in a dataset. Earlier works include [9], [18] and [21] ; more recent work has been done on using resampling-based methods (e.g., see [5],[12] and [26]). In addition, there have been alternative clustering methodologies developed for microarray data (e.g. [5],[15]).

We develop a new approach to finding disease subtypes based on mixture models. A major challenge in this setting is the number of samples profiled in a study (between 20 to 80) is larger than the number of genes per chip (order of several thousands). This problem was recognized in [16] for which a two-stage method was suggested. First, the data are projected into a lower-dimensional subspace using singular value decomposition (SVD) or principal components. The second stage involves fitting a normal mixture model was fit to the projected data by maximum likelihood estimation. Based on the fitted mixture model, the number of clusters can be selected using a Bayesian-based information criterion, and conditional on this quantity, samples could be assigned into each of the groups.

A technique that has emerged recently as a competitor to SVD analysis in the machine learning literature is random projections [10]. This technique has been applied to the motif-finding problem in computational biology [8]. In this article, we combine the mixture model methodology for microarray data with random projection ideas. An attractive feature of the model-based approach is that it provides a probabilistic criterion for assessing the number of true clusters in a given dataset.

There has been other work on applying mixture models to clustering gene expression data. In [30] mixture models for the analysis of microarray data were studied; the focus there was on clustering genes, while we wish to cluster samples. Clustering samples is of great interest in gene expression profiling studies. The classical method for grouping

tumors by doctors has been based on histopathological evaluation of the samples. Despite this taxonomy, there is still great heterogeneity in terms of clinical outcome among tumors classified to be within the same group. High-throughput technologies, such as microarrays, offer the potential of developing a finer classification profile based on gene expression that will lead to better prediction of clinical response than the classical staging method. In addition, such classification systems might also lead to the potential for more individually tailored treatment regimens.

Two papers [4,19] have proposed likelihood-based methods that iterate between clustering the gene expression data and looking for common regulatory motifs in the upstream regions of genes that are coordinately expressed. Our proposal differs from these in several respects. First, we do not consider the gene sequence data because looking for common regulatory motifs is problematic due to the presence of exons. Second, we consider more general covariance structures for the clusters; we will describe them later in the paper. In addition, these authors used mixture models with normally distributed components. Our experience has been that in microarray experiments, there tend to be many genes that are outliers. This has been also noted by other authors [17,24]. Consequently, we will develop robust estimation for mixture models based on the multivariate t distribution [22] and consider performing inference on the number of components using an approximate Bayesian information criterion.

The format of the paper is as follows. In Section 2, we describe the data and model and provide some results for random projections for mixtures of multivariate t -distributions. Section 3 features our proposal for estimation of the parameters using random projections and the expectation-maximization algorithm. We next demonstrate applications of our methodology to simulated and real microarray data. Our main comparisons will be with the two-stage clustering procedure of [16] and with a similar procedure using normal mixture models instead of t -distribution mixture models [10]. Some concluding remarks are made in Section 5.

2. Mixture models

2.1 Data

Let $\mathbf{y} \equiv (\mathbf{y}_1, \dots, \mathbf{y}_n)$ denote the gene expression profiles to be clustered, where $\mathbf{y}_i$ is a p -dimensional vector ($i = 1, \dots, n$). We will assume that the data have already been preprocessed and normalized both within and across slides.

We formulate the following model: observations are independent and identically

distributed realizations from the density

$$f(\mathbf{y}_1, \dots, \mathbf{y}_n) \equiv \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i | \nu, \mu_k, \Sigma_k), \quad (1)$$

where π_k ($k = 1, \dots, K$) is the probability that an observation belongs to the k th group, and

$$f_k(\mathbf{y}_i | \nu, \mu_k, \Sigma_k) \equiv \frac{|\Sigma_k|^{-1/2} \Gamma\{(\nu + p)/2\}}{\Gamma(1/2)^p \Gamma(p/2) \nu^{p/2}} \times \left\{ 1 + \frac{(\mathbf{y}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{y}_i - \mu_k)}{\nu} \right\}^{-(\nu+p)/2}, \quad (2)$$

is a multivariate t density with mean μ_k , covariance matrix Σ_k and degrees of freedom ν . Thus in (1), we have represented the gene expression data as arising from a mixture of K multivariate t -distributions.

To better understand the clusters from a geometric perspective, we use the spectral-based parametrization of Σ_k ($k = 1, \dots, K$), similar to that suggested for multivariate normal mixture models in [3]:

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T, \quad (3)$$

where $\mathbf{D}_k$ is the matrix of eigenvectors of Σ_k , λ_k is a constant of proportionality, and $\mathbf{A}_k$ is a diagonal matrix of values proportionate to the eigenvalues of Σ_k . The parameter λ_k in (3) controls the volume for the k th component, $\mathbf{A}_k$ its shape and $\mathbf{D}_k$ its orientation. These characteristics can be estimated from the data. We have listed possible covariance matrix specifications in Table 1.

There are two points to note about our model. First, we have chosen a multivariate t distribution for the component densities instead of the more popular multivariate normal distribution because the data from microarray studies tend to have relatively heavy tails corresponding to the outlying observations in microarray data. Reasons for outliers include experimental variation in microarray technologies as well as biological variation [17,24].

Second, in what follows, we assume the parameter ν to be the same for all K components in the model. While we could generalize (1) to allow for ν_k ($k = 1, \dots, K$) as unknown parameters to be estimated, we choose not to do so for several reasons. First, this will simplify the computational algorithm necessary for estimation of the parameters in the mixture model. Second, there is no biological reason *a priori* to believe that different groupings of samples should have different degrees of freedom. Instead it is the location and scale parameters (μ_k and Σ_k) that we expect to vary between clusters. Third, based on the model selection criterion we derive later, we can

test to see if different values for ν yield better fits to the model. Our main goals are to make inference on K , the number of components and conditional on K , to assign samples into each of the subgroups.

2.2 Random projections for mixtures of t distributions

As mentioned before, in gene expression studies involving microarray data, p will be typically larger than n . In order to reduce the dimensionality of p , we use random projection methods. While the relevant theory has been developed for random projections based on multivariate normal mixtures [10], it has not been developed for multivariate t -distribution mixture models. However, since the methods of proof for the results presented here are quite similar to those in [10], we state our results as lemmas and refer the interested reader to [10] for the details. Let us first consider the case where $K = 1$ and $\Sigma = \sigma^2 \mathbf{I}_p$. Then the norm of one observation $\mathbf{X}$ has expected squared Euclidean norm $\|\mathbf{X}\|^2 \equiv n\sigma^2\nu(\nu - 2)^{-1}$. Based on arguments using the moment generating function, we can derive the following result.

Lemma 1: For any $\epsilon > 0$, there exist constants $C_1, C_2 > 0$ such that

$$P(|\|\mathbf{X}\|^2 - n\sigma^2\nu(\nu - 2)^{-1}| > n\sigma^2\nu(\nu - 2)^{-1}\epsilon) \leq C_1 \exp(-C_2 n\epsilon^2).$$

The result of Lemma 1 is the mass of most of the distribution of $\mathbf{X}$ is concentrated around a region of radius $\{n\sigma^2\nu(\nu - 2)^{-1}\}^{1/2}$. For more general covariance matrices Σ , the same type of result will hold: namely, the majority of the mass of the distribution will lie within an ellipsoid of radius $\|\mathbf{x}\|_{\Sigma} = n^{1/2}$, where $\|\mathbf{x}\|_{\Sigma} = \mathbf{x}^T \Sigma^{-1} \mathbf{x}$.

Before stating the results involving random projections, we first define separability of t -distributions. Two multivariate t distributions $t(\mu_1, \Sigma_1, \nu)$ and $t(\mu_2, \Sigma_2, \nu)$ are said to be c -separated if $\|\mu_1 - \mu_2\| \geq c\nu(\nu - 2)^{-1} \max(\lambda_{\max}(\Sigma_1), \lambda_{\max}(\Sigma_2))$, where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of the matrix Σ . A mixture of t -distributions is c -separated if its component distributions are pairwise c -separated. Based on these definitions, we state the next result.

Lemma 2: Suppose that the distribution in (1) is c -separated. Let $\delta > 0$ and $\epsilon > 0$. Then with probability at least $1 - \delta$, there exists an $M > 0$ such that the projection of the mixture of t distributions onto a random d -dimensional subspace yields a $(c\sqrt{1 - \epsilon})$ -separated mixture of t distributions in R^d , provided that $d \geq M\epsilon^{-2} \log(k^2/\delta)$.

While we have not explicitly defined the random projection, it is a linear mapping from R^p to R^d . Thus, if the original data are from a mixture distribution of t distributions, so too are the projected data. If the original data are separated, then by Theorem 2, projecting the data into a lower-dimensional subspace will preserve the separation with fairly high probability. In fact, by Theorem 2, we can project a mixture of k multivariate t -distributions of dimension p into a space of dimension $d = O(\log k)$ and still retain most of the separation between the mixture components.

The other result involves the eccentricity associated with the covariance matrices of the component distributions in (3). The eccentricity of a matrix Σ is the ratio of its largest eigenvalue to its smallest eigenvalue.

Lemma 3: For any $0 < \epsilon \leq 1$ and $\delta > 0$, if the high-dimensional eccentricity of the k -component mixture model (3) is at most $O(n^{1/2}\epsilon\delta/\log k)$, then with probability at least $1 - \delta$, the projected t components will have eccentricity less than or equal to 2.

The conclusion of Lemma 3 is that even if the original data are high-dimensional, the projection of the data into the lower-dimensional subspace leads to more spherical clusters. This type of result is similar to results for projection pursuit methods [11]. As mentioned before, Lemmas 1 – 3 can be proven using the techniques in [10].

3. Estimation Procedure and Algorithms

The results from the previous section motivate the following estimation procedure for the mixture model for clustering samples based on microarray data.

1. Randomly project each p -dimensional observation into a d -dimensional subspace, where $d < p$.
2. Fit a mixture model based on the t -distribution using maximum likelihood estimation.
3. Assign the original observations the labels based on the mixture model results in step 2.

Recall that our goal is to determine if there are any disease subtypes in the data. The purpose in formulating model (1) is to ultimately make inference about K , the number of subgroups in the data, and to assign samples to each of the subgroups. Before addressing this issue, we examine steps 1 and 2 of the proposed algorithm in more detail.

3.1 Generating random projections

A random projection can be represented as a d by p matrix. Since it does not depend on the data, it is easily generated using the following approach:

1. Generate a $d \times p$ matrix of iid $N(0,1)$ random variables.
2. Make the d rows of the matrix orthogonal by using the Gram-Schmidt algorithm.

This algorithm takes $O(d^2p)$ operations and results in a $d \times p$ matrix $\mathbf{A}$. One parameter in this procedure that needs to be pre-specified by the user is d . While it might be possible to make d a parameter to be estimated or optimized in some fashion, our approach has been to try several values of d in order to assess sensitivity of the clustering results.

3.2 Maximum Likelihood Estimation using the EM algorithm

Suppose that ν is known in (1). After having applied the projection matrix $\mathbf{A}$ to the original data, we have iid observations $\tilde{\mathbf{y}}_i \equiv \mathbf{A}\mathbf{y}_i$ ($i = 1, \dots, n$) from the density

$$\tilde{f}(\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_n) \equiv \prod_{i=1}^n \sum_{k=1}^K \pi_k \tilde{f}_k(\tilde{\mathbf{y}}_i | \nu, \tilde{\mu}_k, \tilde{\Sigma}), \quad (4)$$

where

$$\tilde{f}_k(\tilde{\mathbf{y}}_i | \nu, \tilde{\mu}_k, \tilde{\Sigma}_k) \equiv \frac{|\tilde{\Sigma}_k|^{-1/2} \Gamma\{(\nu + d)/2\}}{\Gamma(1/2)^d \Gamma(d/2) \nu^{d/2}} \times \left(1 + \frac{(\tilde{\mathbf{y}}_i - \tilde{\mu}_k)^T \tilde{\Sigma}_k^{-1} (\tilde{\mathbf{y}}_i - \tilde{\mu}_k)}{\nu_k} \right)^{-(\nu+d)/2}, \quad (5)$$

$\tilde{\mu}_k = \mathbf{A}\mu_k$, $\tilde{\Sigma}_k = \mathbf{A}\Sigma_k\mathbf{A}^T$. By a basic property of the multivariate t-distribution,

$$\tilde{f}_k(\tilde{\mathbf{y}}_i | \nu, \tilde{\mu}_k, \tilde{\Sigma}_k) = \phi(\tilde{\mathbf{y}}_i | \tilde{\mu}_k, u_i^{-1} \tilde{\Sigma}_k),$$

where $\phi(\cdot | \mu, \Sigma)$ is the pdf of a multivariate normal distribution with mean μ and covariance matrix Σ , and u_i are iid realizations from a $\chi^2_{\nu+1}/(\nu+1)$ random variable.

To maximize the likelihood in (4), we will use the expectation maximization (EM) algorithm. In the EM algorithm framework, the complete data are $\mathbf{x}_i = (\tilde{\mathbf{y}}_i, \mathbf{z}_i, u_i)$ ($i = 1, \dots, n$), where $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ is defined by

$$z_{ik} \equiv \begin{cases} 1 & \text{if } \tilde{\mathbf{y}}_i \text{ belongs to group } k \\ 0 & \text{otherwise,} \end{cases}$$

and $u_1, \dots, u_n$ are realizations from $U_1, \dots, U_n$, where $U_i|z_{ik} = 1$ are independent observations from a $\chi_\nu^2/(\nu + 1)$ random variable. It is easy to then derive the complete-data log likelihood:

$$l(\tilde{\theta}_1, \dots, \tilde{\theta}_K, \pi_1, \dots, \pi_K | \mathbf{x}_1, \dots, \mathbf{x}_n) = \log L_1(\pi) + \log L_2(\nu) + \log L_3(\tilde{\theta}_1, \dots, \tilde{\theta}_K),$$

where

$$\log L_1(\pi) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log \pi_k,$$

$$\log L_2(\nu) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \{-\log \Gamma(\nu/2) + \nu/2 \log(\nu/2) + \nu/2(\log u_i - u_i) - \log u_i\},$$

$$\log L_3(\tilde{\theta}_1, \dots, \tilde{\theta}_K) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \{-(d/2) \log(2\pi) - (1/2) \log |\tilde{\Sigma}_k| - u_i/2 (\tilde{\mathbf{y}}_i - \tilde{\mu}_k)^T \tilde{\Sigma}^{-1} (\tilde{\mathbf{y}}_i - \tilde{\mu}_k)\}$$

and $\tilde{\theta}_k = (\tilde{\mu}_k, \tilde{\Sigma}_k)$. Based on the complete data log-likelihood, the EM algorithm works by maximizing the expected value of the complete data log-likelihood, conditional on observed data:

$$Q(\Psi; \Psi^{(m)}) \equiv E[l(\Psi | \mathbf{x}_1, \dots, \mathbf{x}_n) | \tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_n, \Psi^{(m)}],$$

where $\Psi = (\pi_1, \dots, \pi_K, \tilde{\theta}_1, \dots, \tilde{\theta}_K)$, and $\Psi^{(m)}$ is the value of Ψ at the m th iteration.

The E-step of the algorithm at the $(k+1)$ th step involves computing the expected value of the complete data log-likelihood conditional on the observed data; this involves computing $E(Z_{ik} | \tilde{\mathbf{y}}_i, \Psi^{(m)})$, $E(U_j | \tilde{\mathbf{y}}_i, \mathbf{z}_i, \Psi^{(m)})$ and $E(\log U_j | \tilde{\mathbf{y}}_i, \mathbf{z}_i, \Psi^{(m)})$. By Bayes' rule, we have that

$$E(Z_{ik} | \tilde{\mathbf{y}}_i, \Psi^{(m)}) = \frac{\pi_k^{(m)} \tilde{f}_k(\tilde{\mathbf{y}}_i | \nu, \tilde{\theta}_k^{(m)})}{\tilde{f}(\tilde{\mathbf{y}}_i)},$$

where $\pi_k^{(m)} = n^{-1} \sum_{i=1}^n E(Z_{ik} | \tilde{\mathbf{y}}_i, \Psi^{(m)})$. Using arguments from [23], we have that

$$E(U_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i, \Psi^{(m)}) \equiv \frac{\nu + p}{\nu + \delta(\tilde{\mathbf{y}}_i; \tilde{\theta}_k^{(m)})}$$

and

$$E(\log U_i | \tilde{\mathbf{y}}_i, \mathbf{z}_i, \Psi^{(m)}) \equiv \psi\left(\frac{\nu + p}{2}\right) - \log \left[\frac{1}{2} \{\nu + \delta(\tilde{\mathbf{y}}_i; \tilde{\theta}_k^{(m)})\} \right],$$

where $\delta(\mathbf{y}; \tilde{\theta}) = (\mathbf{y} - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\mathbf{y} - \tilde{\mu})$, and $\psi(s)$ is the Digamma function.

The M-step of the EM-step involves maximizing $Q(\Psi; \Psi^{(m)})$ with respect to Ψ . The estimators of $\tilde{\theta}_k$ at the $(m+1)$ th iteration are given by

$$\tilde{\mu}_k^{(m+1)} = \sum_{i=1}^n \tau_{ik}^{(m+1)} u_{ij}^{(m+1)} \tilde{\mathbf{y}}_i / \sum_{i=1}^n \tau_{ik}^{(m+1)} u_{ij}^{(m+1)},$$

and

$$\tilde{\Sigma}_k^{(m+1)} = \sum_{i=1}^n \tau_{ik}^{(m+1)} u_{ik}^{(m+1)} (\tilde{\mathbf{y}}_i - \tilde{\mu}_k^{(m+1)}) (\tilde{\mathbf{y}}_i - \tilde{\mu}_k^{(m+1)})^T / \sum_{i=1}^n \tau_{ik}^{(m+1)} u_{ij}^{(m+1)}.$$

In the EM algorithm described above, it was assumed that ν was fixed. We estimate ν by grid search, using the likelihood as the objective function to be maximized. One of the determinants of the convergence of the EM algorithm is the choice of starting values. While one possible option is using random initialization, we instead propose a procedure for determining initial values for the EM algorithm.

3.3 Hierarchical agglomerative clustering for initial values

Suppose that for each observation $\mathbf{y}_i$ ($i = 1, \dots, n$), the true cluster membership l_i ($\in \{1, \dots, K\}$) is known. We can then define a classification log-likelihood

$$l^{CL}(\tilde{\theta}_1, \dots, \tilde{\theta}_K; l_1, \dots, l_n | \tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_n) \equiv \prod_{i=1}^n \tilde{f}_{l_i}(\tilde{\mathbf{y}}_i), \quad (6)$$

where $\tilde{f}_{l_i}(\mathbf{y}_i | \theta_{l_i})$ is the density (5) ($i = 1, \dots, n; k = 1, \dots, K$). Potential specifications for the covariance matrices in hierarchical agglomerative clustering can be found in Table 1.

In our situation, however, the true cluster memberships are unknown. We use agglomerative hierarchical clustering in order to find a maximizer of (6). This procedure is similar in spirit to other hierarchical clustering algorithms; it starts with each observation as a singleton cluster. All possible pairs of clusters are considered, and the pair that leads to the greatest increase in (6) is then merged. This procedure is iterated repeatedly until one cluster remains. By default, this cluster will have all n observations. Note that with agglomerative hierarchical clustering, the clusters are merged according to the probabilistic criterion (6). This is different from ordinary hierarchical clustering methods, where the merging of clusters is based on a combination of the dissimilarity matrix and a method of defining distance between clusters (e.g., average linkage).

In Section 3.2, we left ν unknown. For the agglomerative hierarchical clustering procedure described here, we will fix $\nu = 3$. This represents the least amount of information possible for each cluster.

3.4 Choosing the number of clusters

We now derive a Bayesian-based information criterion (BIC) for selecting the number of clusters, related to the method of Bayes factors as described in [23]. Suppose we fix

the number of components in the mixture model to be K_1 . Suppressing the dependence on K_1 , if we let $f(\Psi)$ denote the prior density for Ψ , then the integrated likelihood is given by

$$f(\tilde{\mathbf{y}}) = \int \exp\{\log f(\Psi, \tilde{\mathbf{y}})\} d\Psi, \quad (7)$$

where $f(\Psi, \tilde{\mathbf{y}}) = f(\Psi)L(\Psi|\tilde{\mathbf{y}})$ and $L(\Psi|\tilde{\mathbf{y}}) = \tilde{f}(\tilde{\mathbf{y}})$. Taking a second order Taylor's series expansion of $\log f(\Psi, \tilde{\mathbf{y}})$ about the posterior mode $\tilde{\Psi}$, we have that

$$\log f(\Psi, \tilde{\mathbf{y}}) = \log f(\tilde{\Psi}, \tilde{\mathbf{y}}) + \partial \log f(\tilde{\Psi}, \tilde{\mathbf{y}})/\partial \Psi - \frac{1}{2}(\Psi - \tilde{\Psi})^T \mathbf{H}(\tilde{\Psi})(\Psi - \tilde{\Psi}) + o_P(n^{-1/2}), \quad (8)$$

where $\mathbf{H}(\Psi)$ is the negative Hessian matrix of $\log f(\Psi, \tilde{\mathbf{y}})$. Because $\partial \log f(\tilde{\Psi}, \tilde{\mathbf{y}})/\partial \Psi = 0$ the second term in (8) vanishes. Plugging this into the integral (7) yields

$$f(\tilde{\mathbf{y}}) = f(\tilde{\Psi}, \tilde{\mathbf{y}})(2\pi)^{d/2}|\mathbf{H}(\tilde{\Psi})|^{-1/2}. \quad (9)$$

Taking logarithms of both sides gives

$$\log f(\tilde{\mathbf{y}}) \approx \log L(\tilde{\Psi}) + \log p(\tilde{\Psi}) - \frac{1}{2} \log |\mathbf{H}(\tilde{\Psi})| + \frac{d}{2} \log(2\pi).$$

If we can assume that all models are equal likely, then we can write

$$\log f(\tilde{\mathbf{y}}) \approx \log L(\tilde{\Psi}) + \log p(\tilde{\Psi}) - \frac{1}{2} \log |\mathbf{I}(\tilde{\Psi}; \tilde{\mathbf{y}})| + \frac{d}{2} \log(2\pi), \quad (10)$$

where $\mathbf{I}(\tilde{\Psi}; \tilde{\mathbf{y}})$ is the observed information matrix for the posterior density. Ignoring $O(1)$ terms in (10) and observing that $|\mathbf{I}(\tilde{\Psi}; \tilde{\mathbf{y}})| = O(d \log n)$ yields a model selection criterion $-2 \log L(\tilde{\Psi}) + d \log n$ for comparing mixture models of differing numbers of components. We can also use this quantity for determining the appropriate degrees of freedom ν , but that will not be pursued here.

Conditional on the estimated value of K , we can assign observations into each of the K subgroups based on the posterior probabilities. Thus, the mixture model approach avoids the greedy assignment of samples into clusters as in hierarchical clustering methods.

The suitability of BIC was first addressed in [30], where they found it to be a successful method in determining the number of clusters in terms of cluster validity.

4. Results

4.1 Simulated data

We report the results from two sets of simulation studies. In the first simulation study, the goal was to compare the classification accuracy for detecting subtypes between the method proposed here with that given in [16]. Because we are dealing with simulated data, we know the labels and hence can determine which subtypes are correctly classified. We generated random vectors from (1) with $K = 2$ subtypes of dimension $p = 1000$ and $n = 100$ samples. For the first subtype, the first 50 genes were overexpressed by a $\log(6)$ -fold change relative to the second. We chose $\Sigma = \lambda \mathbf{I}_p$ with $\lambda = 0.15$. We also varied the number of components used in the SVD analysis and d in the random projections. For each simulation setting, we performed 200 simulation samples. The results are summarized in Table 2. Based on this table, we find that the classification accuracy for the random projection method tends to be 10-15% higher than for the two-stage SVD method of [16]. Also, the classification accuracies do not appear to depend on d for this scenario.

The other simulation involved determining K , the number of subtypes in the data. We generated the data as before and compared selection based on the approximate Bayes factors developed in Section 3.4 to the proposals of [9],[18] and [21] based on a dendrogram using average linkage hierarchical clustering. We looked at the estimators of K under these methods with that selected by the Bayes factor. We took the true value of K to be 1,2,3 and 4 in the simulations. The results are given in Table 3. While the Bayes factor procedure outperforms the other methods for $K = 2, 3, 4$, its greatest usefulness appears to be in discerning situations in which there are no clusters (i.e. $K = 1$). The methods in [9] and [21] cannot detect this situation, while the approach of [18] has a low percentage of classifying this situation correctly.

4.2 Melanoma data

In this section, we consider data from the cDNA microarray experiments described in [7]. For the analysis presented here, we consider the 31 melanoma samples. The microarray used in this study contained 8,150 human cDNAs, of which 6,912 were sequence verified. Of these cDNAs, 3,613 were included in the analysis. There was a preprocessing of the data, which is described in [16]. No normalization was performed across experiments, since a single reference probe was used for all experiments. In [7], an average-linkage hierarchical clustering procedure for clustering samples. A replica of their dendrogram can be found in Figure 3. Based on a dendrogram cutoff value of 0.54, they found a cluster of 19 melanomas. They assessed the reproducibility of this cluster using some simulation-based tests. As reported in [16], the methods of [9] and [18] yield a value of 2 clusters, implying that the dendrogram in Figure 1 should be cut

at a value between 1.05 and 1.15. The method of [21] yields a value of eight clusters, implying the same cutoff value as that in [7].

We first demonstrate the method of [16]. A singular value decomposition was performed. This yielded the principal components, whose variances are given in Figure 1. We then plotted the individual principal components to see if a mixture model seemed reasonable. A plot for the first principal component is given in Figure 2. A mixture model seems to be adequate for these data. By graphical inspection, we considered the first and third principal components. We fit mixture models to these two components based on mixture models from the multivariate t distribution. The results are summarized in Table 3. None of the final models validate the cluster of melanomas found in [7]. The closest possible validation occurs when $\Sigma_k = \lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}^T$. This analysis yields a cluster of 17 melanomas that include samples such as W1791-C, UACC-647, and UACC-1012, which were not included in the original Bittner et al. cluster. The 17-melanoma cluster and their inclusion in the Bittner 19-melanoma cluster are listed in Table 5.

The cancer samples are now clustered using the random projection methodology described in the paper. We first project the data into $d = 2$ dimensions and then fit a mixture model based on the bivariate t distribution. The resulting analyses are also presented in Table 3. Based on these results, we find that the Bayes factor approach tends to select either 5 clusters or 10 clusters. The clusters are provided in Table 6. However, the analyses lead to groupings that are different from that suggested in [7].

One potential method for validating the clustering results found is to see if they are predictive of clinical outcome. We examine that here by studying the relationship between the clustering patterns with survival. A caveat of the analysis is that survival data are only publicly available on 15 out of the 31 samples; these are provided in the supplementary information for [7]. For the five groups found using the proposed method, only two clusters had sufficiently large number of observations with survival information. The plot of survival in the two groups is given in Figure 4. For the sake of numerical comparison, we fit a normal mixture model to the data after random projections and assigned the individual to two groups. As mentioned by a referee, the normal mixture model in principle should work as well as the proposed method. The estimated survival curves for the two groups are given in Figure 5. Comparing the two plots, we find less separation between the groups in Figure 5 relative to Figure 4. If we calculate a log-rank test for separation, we get a value of 3.8 for Figure and 1.44 for Figure 5. It should be noted that larger values of the log-rank statistic correspond to greater separation.

5. Discussion

The generation of large datasets from gene expression experiments in cancer profiling studies necessitates the need for fast scalable algorithms for determining disease subtypes. At the same time, by having methods that are probabilistic, we can model the inherent uncertainty and variation in the gene expression data [2]. The fact that the dimension of the gene expression profile is larger than the the number of samples profiled is a novel aspect to the problem. An approach for clustering of the cancer samples using a combination of random projections and mixture models has been proposed in this paper. We have chosen to model the gene expression patterns using a mixture of multivariate t distributions instead of the more popular multivariate normal mixture model because in our experience, we have found many outlying observations.

The primary advantage of the proposed method using random projection is that it will quite scalable for larger datasets. The singular value decomposition will in general require $O(p^3)$ operations, which will be computationally intensive when p , the number of genes, is large. By contrast, the number of operations for the proposed method is simply linear in the number of genes. The theoretical results in Section 2.2, which can be proven by techniques in [10], imply that fitting mixture models to the projected data is a sensible thing to do.

While we developed estimation methods for the parameters in the mixture models based on the EM algorithm, the model selection criteria that were developed were Bayesian in nature. An alternative approach to modelling the gene expression data in a “large p , small n ” setting is to use fully Bayesian methods for both inference and model selection. Some work on this has been done for the classification of microarray data [28]. The disadvantages of this approach are this they tend to be computationally intensive. However, it would be important to contrast the results of this study with those from a Bayesian approach. One problem is the formulation the probability model corresponding to the projection. While such work has been done for singular value decomposition analysis [25], we know of no such work for random projections.

Acknowledgments

This work has been supported by a MUNN Idea Grant and Prostate SPORE Seed Grant from the University of Michigan, as well as a Bioinformatics Pilot Award from the University of Michigan and Pfizer.

References

- [1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt, *Nature* 403 (2000), 503–511.
- [2] P. B. Baldi and S. Brunak, *Bioinformatics: A Machine Learning Approach* (MIT Press, Cambridge, MA, 2002).
- [3] J. D. Banfield and A. E. Raftery, *Biometrics* 49 (1993), 803–821.
- [4] Y. Barash and N. Friedman, *Journ. Comput. Biol.* 9 (2002), 169 – 191.
- [5] A. Ben-Dor, R. Shamir, and Z. Yakhini, *Journ. Comput. Biol.* 6 (1999), 281–297.
- [6] A. Ben-Hur, A. Elisseeff, and I. Guyon, A stability-based method for discovering structure in clustered data, in R. B. Altman et al., ed., *Proceedings of Pacific Symposium on Biocomputing 2002* (World Scientific Press, New Jersey, 2002), 6 – 17.
- [7] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts and V. Sondak, *Nature* 406 (2000), 536–540.
- [8] J. Buhler, *Bioinformatics* 17 (2001), 419–428.
- [9] R. B. Calinski and J. Harabasz, *Comm. in Stat.* 3 (1974), 1–27.
- [10] S. Dasgupta (2000). *Learning probability distributions*, Ph.D. dissertation, Department of Computer Science, UC-Berkeley.
- [11] P. Diaconis and D. Freedman, *Annals of Statistics* 12 (1984), 793–815.
- [12] S. Dudoit and J. Fridlyand, *Genome Biology* 3 (2002), RESEARCH0036.
- [13] B. Efron, R. Tibshirani, J. D. Storey and V. Tusher, *Journ. Am. Statist. Assoc.* 96 (2001), 1151 – 1160.

- [14] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, *Proc. Nat. Acad. Sci.* 95 (1998), 14863–14868.
- [15] G. Getz, E. Levine, and E. Domany, *Proc. Nat. Acad. Sci.* 97 (2000), 12079 – 12084.
- [16] D. Ghosh and A. M. Chinnaiyan, *Bioinformatics* 18 (2002), 275 – 286.
- [17] A. B. Goryachev, P. F. Macgregor and A. M. Edwards, *J. Comput. Biol.* 8, 443 – 461.
- [18] J. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975).
- [19] I. Holmes and W. J. Bruno, Finding regulatory elements using joint likelihoods for sequence and expression profile data, in: R. Altman et al., Eds., *Proceedings Eighth Annual International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, La Jolla, CA, 2000), 202–210.
- [20] J. G. Ibrahim, M.-H. Chen and R. J. Gray, *Journ. Am. Statist. Assoc.* 97 (2002), 88 – 99.
- [21] W. J. Krzanowski and Y. T. Lai, *Biometrics* 44 (1985), 23–34.
- [22] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, *Journ. Am. Statist. Assoc.* 84 (1989), 881 – 896.
- [23] G. McLachlan and D. Peel, *Finite Mixture Models.* (Wiley, New York, 2000).
- [24] R. Sasik, E. Calvo and J. Corbeil, *Bioinformatics* 18 (2002), 1633 – 1640.
- [25] M. E. Tipping and C. M. Bishop, *Journ. Roy. Statist. Soc. Ser. B* 61 (1999), 611 – 622.
- [26] R. Tibshirani, G. Walter and T. Hastie, *Journ. Roy. Statist. Soc. Ser. B* 63 (2001), 411 – 423.
- [27] G. C. Tseng, M.-K. Oh, L. Rohlin, J. C. Liao and W. H. Wong, *Nucleic Acids Res.* 29 (2001), 2549–2557.
- [28] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, *Proc. Nat. Acad. Sci.* 98 (2001), 11462 – 11467.

- [29] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, Nucleic Acids Research 30 (2002), e15.
- [30] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo, Bioinformatics 17 (2001), 977–987.

Table 1. Possible covariance structures for Σ_k ($k = 1, \dots, K$) for agglomerative hierarchical clustering (HC) and EM algorithm (EM).

Variance Model	HC	EM	Distribution	Volume	Shape	Orientation
$\lambda \mathbf{I}$	x	x	Spherical	equal	equal	NA
$\lambda_k \mathbf{I}$	x	x	Spherical	variable	equal	NA
$\lambda \mathbf{D} \mathbf{A} \mathbf{D}^T$	x	x	Ellipsoidal	equal	equal	equal
$\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$		x	Ellipsoidal	equal	equal	variable
$\lambda \mathbf{D}_k \hat{\mathbf{A}} \mathbf{D}_k^T$	x		Ellipsoidal	equal	fixed	variable
$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$		x	Ellipsoidal	variable	equal	variable
$\lambda_k \mathbf{D}_k \hat{\mathbf{A}} \mathbf{D}_k^T$	x		Ellipsoidal	variable	fixed	variable
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	x	x	Ellipsoidal	variable	variable	variable

Table 2. Simulation results for comparing classification accuracy of random projections-based method with SVD-based method of Ghosh and Chinnaiyan (2002).

Σ_k	Random projections			SVD method		
	$d = 2$	$d = 3$	$d = 4$	$r = 2$	$r = 3$	$r = 4$
$\lambda \mathbf{I}$	0.73	0.72	0.71	0.60	0.62	0.61
$\lambda_k \mathbf{I}$	0.71	0.70	0.71	0.61	0.63	0.61
$\lambda \mathbf{DAD}^T$	0.71	0.71	0.71	0.62	0.62	0.61
$\lambda \mathbf{D}_k \hat{\mathbf{A}} \mathbf{D}_k^T$	0.72	0.71	0.72	0.61	0.60	0.60
$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	0.73	0.71	0.72	0.61	0.61	0.61
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	0.73	0.73	0.71	0.62	0.61	0.61

Note : First r principal components used in Ghosh and Chinnaiyan procedure; Σ_k represents covariance structure used to fit t mixture model to projected components.

Table 3. Simulation results for comparing Bayes factor selection procedure for choosing number of clusters to other methods.

K	Random projections			CH	KL	H
	$d = 2$	$d = 3$	$d = 4$			
1	1.00	0.99	0.99	-	-	0.28
2	0.96	0.95	0.94	0.88	0.87	0.88
3	0.93	0.94	0.93	0.85	0.85	0.84
4	0.95	0.93	0.93	0.83	0.82	0.81

Note : CH: method of Calinski and Harabasz (1974); KL: method of Krzanowski and Lai (1985); H: method of Hartigan (1975). Numbers in table represent proportion of times K clusters were selected by procedure out of 100 trials.

Table 4. Results for melanoma data using random projections-based method and SVD-based method of Ghosh and Chinnaiyan (2002).

Σ_k	Random projections	SVD method
$\lambda \mathbf{I}$	10	10
$\lambda_k \mathbf{I}$	5	3
$\lambda \mathbf{D} \mathbf{A} \mathbf{D}^T$	5	5
$\lambda \mathbf{D}_k \hat{\mathbf{A}} \mathbf{D}_k^T$	9	4
$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	9	2
$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$	5	4

Note : First r principal components used in Ghosh and Chinnaiyan procedure; Σ_k represents covariance structure used to fit t mixture model to projected components.

Table 5. Cluster of 17 melanomas found by Ghosh and Chinnaiyan procedure

TC-F027*	UACC-1012	UACC-647	UACC-827	UACC-1097	TD-1720*
TD-1638*	TD-1730*	TD-1376-3*	TC-1376-3*	UACC-1022*	UACC-3093*
UACC-457*	UACC-2534*	UACC-1256*	UACC-091*	M93-007*	

Note : Asterisk indicates that the sample was in 19 melanoma cluster found by Bittner et al. (2000)

Table 6. Clusters of melanomas using random projections with $\Sigma_k = \lambda \mathbf{D} \hat{\mathbf{A}}_k \mathbf{D}^T$
 $(k = 1, \dots, 5)$

Cluster	1	2	3	4	5
	UACC-647	UACC-3093	TC-F027	UACC-827	UACC-2873
	WM1791-C	M92-001	UACC-1012	HA-A	TD-1376-3
	UACC-903		UACC-1529	TD-1384	TC-1376-3
	TD-1720		UACC-930	UACC-1022	UACC-383
	M91-054		UACC-1097	UACC-457	UACC-1273
			M93-47	UACC-2534	UACC-091
			TD-1638	UACC-1256	M93-007
			TD-1730		
			A-375		
			UACC-502		

Note : $K = 5$ determined using Bayesian criterion.

Figure 1: Graph of variances for first 10 principal components for melanoma data.

Figure 2: Probability histogram for first principal component for melanoma data from Bittner et al.

Figure 3: Hierarchical Clustering Dendrogram of gene expression data from Bittner et al. (2000). Average linkage clustering used.

Figure 4: Survival experience for clusters 3 and 5 from Table 6 for Bittner et al. (2000) data. Solid line represents cluster 3, while dashed line is for cluster 5.

Figure 5: Survival experience using random projections and normal mixture model for Bittner et al. (2000) data. Solid line represents cluster 1, while dashed line is for cluster 2.