Model-Based Clustering and Data Transformations for Gene Expression Data

Ka Yee Yeung [1], Chris Fraley [2], Alejandro Murua [3], Adrian E. Raftery [4], Walter L. Ruzzo [5]

Technical Report no. 396
Department of Statistics
University of Washington

April 30, 2001

[1]Dept of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-2350, USA. Email: kayee@cs.washington.edu
[2]Dept of Statistics, University of Washington, Seattle, Washington 98195-4322, USA. Email: fraley@stat.washington.edu
[3]Insightful Corporation, 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA. Email: amurua@insightful.com
[4]Dept of Statistics, University of Washington, Seattle, Washington 98195-4322, USA. Email: raftery@stat.washington.edu
[5]Dept of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-2350, USA. Email: ruzzo@cs.washington.edu

Abstract

Clustering is a useful exploratory technique for the analysis of gene expression data, and many different heuristic clustering algorithms have been proposed in this context. Clustering algorithms based on probability models offer a principled alternative to heuristic algorithms. Model-based clustering assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. This Gaussian mixture model has been shown to be a powerful tool for many applications. In addition, the issues of selecting a "good" clustering method and determining the "correct" number of clusters are reduced to model selection problems in the probability framework.

We benchmarked the performance of model-based clustering on several synthetic and real gene expression data sets for which external evaluation criteria were available. The model-based approach has superior performance on our synthetic data sets, consistently selecting the correct model and the right number of clusters. On real expression data, the model-based approach produced clusters of quality comparable to a leading heuristic clustering algorithm, but with the key advantage of suggesting the number of clusters and an appropriate model. We also assessed the degree to which these real gene expression data sets fit multivariate Gaussian distributions both before and after subjecting them to commonly used data transformations. Suitably chosen transformations seem to result in reasonable fits.

These analyses were carried out using the MCLUST software available at www.stat.washington.edu/fraley/mclust/soft.shtml.

Contents

List of Tables

List of Figures

1 Introduction

DNA microarrays offer the first great hope of studying the variation of many genes simultaneously (Lander 1999). Large amounts of gene expression data have been generated by researchers. There is a great need to develop analytical methodology to analyze and to exploit the information contained in gene expression data (Lander 1999). Because of the large number of genes and the complexity of biological networks, clustering is a useful exploratory technique for the analysis of gene expression data.

A wide range of clustering algorithms have been proposed to analyze gene expression data, including hierarchical clustering (Eisen, Spellman, Brown, and Botstein 1998), self-organizing maps (Tamayo, Slonim, Mesirov, Zhu, Kitareewan, Dmitrovsky, Lander, and Golub 1999), k-means (Tavazoie, Hughes, Campbell, Cho, and Church 1999), graph-theoretic approaches (for example, Ben-Dor and Yakhini 1999 and Hartuv, Schmitt, Lange, Meirer-Ewert, Lehrach, and Shamir 1999), and support vector machines (Brown, Grundy, Lin, Cristianini, Sugnet, Furey, Ares, and Haussler 2000). Success in applications has been reported for many clustering approaches, but so far no single method has emerged as the method of choice in the gene expression analysis community. Most of the proposed clustering algorithms are largely heuristically motivated, and the issues of determining the "correct" number of clusters and choosing a "good" clustering algorithm are not yet rigorously solved. Eisen, Spellman, Brown, and Botstein (1998) and Tamayo, Slonim, Mesirov, Zhu, Kitareewan, Dmitrovsky, Lander, and Golub (1999) used visual display to determine the number of clusters. Yeung, Haynor, and Ruzzo (2001) suggested clustering the data set leaving out one experiment at a time and then compared the performance of different clustering algorithms using the left-out experiment. The gap statistic (Tibshirani, Walther, and Hastie 2000) estimates the number of clusters by comparing within-cluster dispersion to that of a reference null distribution.

Clustering algorithms based on probability models offer a principled alternative to heuristic-based algorithms. In particular, the model-based approach assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. The Gaussian mixture model has been shown to be a powerful tool for many applications (for example Banfield and Raftery 1993, Celeux and Govaert 1993, McLachlan and Basford 1988). With the underlying probability model, the problems of determining the number of clusters and of choosing an appropriate clustering method become statistical model choice problems (Dasgupta and Raftery 1998, Fraley and Raftery 1998, Fraley and Raftery 2000). This is an advantage over heuristic clustering algorithms, in which there is no established method to determine the number of clusters or the best clustering method. Details of the model-based approach and the model selection methodologies are discussed in Section 2.

Since the model-based approach is based on the assumption that the data are distributed according to a mixture of Gaussian distributions, we assess whether this assumption holds before applying the model-based approach. Moreover, the raw gene expression data do not satisfy the Gaussian mixture assumption as we will see in Section 4. Hence, we explore different transformations of gene expression data sets and assess the extent to which the transformed data sets satisfy the normality assumption.

In Section 6, we show that the existing model-based clustering implementations produce higher quality clustering results than a leading heuristic approach when the data set is appropriately transformed. The existing model-based clustering methods were designed for applications other than gene expression, and yet they perform well in this context. We therefore feel that, with further refinements specifically for the gene expression problem, the model-based approach has the potential to become the approach of choice for clustering gene expression data.

2 The Model-Based Clustering Approach

2.1 The Model-Based Framework

The mixture model assumes that each component (group) of the data is generated by an underlying probability distribution. Suppose the data $\mathbf{y}$ consist of independent multivariate observations $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$. Let G be the number of components in the data. The likelihood for the mixture model is

$$\mathcal{L}_{MIX}(\theta_1, \ldots, \theta_G|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y_i}|\theta_\mathbf{k}), \tag{1}$$

where f_k and $\theta_\mathbf{k}$ are the density and parameters of the kth component in the mixture, and τ_k is the probability that an observation belongs to the kth component ($\tau_k \geq 0$ and $\sum_{k=1}^{G} \tau_k = 1$).

In the Gaussian mixture model, each component k is modeled by the multivariate normal distribution with parameters $\mu_\mathbf{k}$ (mean vector) and Σ_k (covariance matrix):

$$f_k(\mathbf{y_i}|\mu_\mathbf{k}, \Sigma_k) = \frac{\exp\{-\frac{1}{2}(\mathbf{y_i} - \mu_\mathbf{k})^T \Sigma_k^{-1} (\mathbf{y_i} - \mu_\mathbf{k})\}}{\sqrt{\det(2\pi\Sigma_k)}}. \tag{2}$$

Geometric features (shape, volume, orientation) of each component k are determined by the covariance matrix Σ_k. Banfield and Raftery (1993) proposed a general framework for representing the covariance matrix in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{3}$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k, and λ_k is a scalar. The matrix D_k determines the orientation of the component, A_k determines the shape of the component, and λ_k determines its volume.

Allowing some but not all of the parameters in Equation (3) to vary yields an array of models within this general framework. In this paper, we consider five such models, outlined below. Constraining $D_k A_k D_k^T$ to be the identity matrix I corresponds to Gaussian mixtures in which each component is spherically symmetric. The *equal volume spherical* model (denoted by EI), which is parameterized by $\Sigma_k = \lambda I$, represents the most constrained model under this framework, with the smallest number of parameters. The *unequal volume spherical* model (VI), $\Sigma_k = \lambda_k I$, allows the spherical components to have different volumes, determined by a different λ_k for each component k. The *unconstrained* model (VVV) allows all of D_k, A_k and λ_k to vary between components. The unconstrained model has the advantage that it is the most general model, but has the disadvantage that the maximum number of parameters need to be estimated, hence requiring relatively more data points in each component.

There is a range of elliptical models with other constraints and fewer parameters. For example, with the parameterization $\Sigma_k = \lambda DAD^T$, each component is elliptical, but all have equal volume, shape and orientation (denoted by EEE). All of these models are implemented in MCLUST (Fraley and Raftery 1998). Celeux and Govaert (1995) also considered the model in which $\Sigma_k = \lambda_k B_k$, where B_k is a diagonal matrix with $|B_k| = 1$. Geometrically, the diagonal model corresponds to axis-aligned elliptical components. In the experiments reported in this paper, we considered the equal volume spherical (EI), unequal volume spherical (VI), EEE and unconstrained (VVV) models as implemented in MCLUST (Fraley and Raftery 1999), and the diagonal model as implemented by Murua, Tantrum, Stuetzle, and Sieberts (2001).

In both the MCLUST implementation and the diagonal model implementation, the model parameters are estimated by the EM algorithm, in which expectation (E) steps and maximization (M) steps alternate. In the E-step, the probability of each observation belonging to each cluster is estimated conditionally on the current parameter estimates. In the M-step, the model parameters are estimated given the current group membership probabilities. When the EM algorithm converges, each observation is assigned to the group with the maximum conditional probability. The EM algorithm is usually initialized with a model-based hierarchical clustering step (Dasgupta and Raftery 1998, Fraley and Raftery 1998).

The classical iterative k-means clustering algorithm, first proposed as a heuristic clustering algorithm, has been shown to be closely related to model-based clustering using the equal volume spherical model (EI), as computed by the EM algorithm (Celeux and Govaert 1992). K-means has been successfully used for a wide variety of clustering tasks, including clustering of gene expression data. This is not surprising, from the model-based perspective, given k-means' interpretation as an approximate estimation method for a parsimonious model of simple independent Gaussians, a model that arises commonly in many contexts.

Nevertheless, it should also be unsurprising that k-means is *not* the right model in many other circumstances. For example, the unequal volume spherical model (VI) may be more appropriate if some groups of genes are much more tightly co-regulated than others. Similarly, the diagonal model also assumes that experiments are uncorrelated, but allows for unequal variances in different experiments, as might be the case in a stress-response experiment or a tumor/normal comparison, say. We have observed considerable correlation between samples in time-series experiments, coupled with unequal variances. One of the more general elliptical models may better fit the data in these cases. One of the key advantages of the model-based approach is the availability of a variety of models that smoothly interpolate between these scenarios (and others). Of course, there is a tradeoff in that the more general models require more parameters to be estimated, and choice of which model to use may seem *ad hoc*. A second key advantage of model-based clustering is that there is a principled, data-driven way to approach the latter problem. This is the topic of the next subsection.

2.2 Model Selection

Each combination of a different specification of the covariance matrices and a different number of clusters corresponds to a separate probability model. Hence, the probabilistic framework of model-based clustering allows the issues of choosing the best clustering algorithm and the correct number of clusters to be reduced simultaneously to model selection problems. This is important because there is a tradeoff between probability model (and the corresponding clustering method), and number of clusters. For example, if one uses a complex model, a small number of clusters may suffice, whereas if one uses a simple model, one may need a larger number of clusters to fit the data adequately.

Let D be the observed data, and M_1 and M_2 be two different models with parameters θ_1 and θ_2 respectively. The *integrated likelihood* is defined as $p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$ where $k = 1, 2$ and $p(\theta_k|M_k)$ is the prior distribution of θ_k. The integrated likelihood represents the probability that data D is observed given that the underlying model is M_k. The Bayes factor (Kass and Raftery 1995) is defined as the ratio of the integrated likelihoods of the two models, *i.e.,* $B_{12} = p(D|M_1)/p(D|M_2)$. In other words, the Bayes factor B_{12} represents the posterior odds that the data were distributed according to model M_1 against model M_2 assuming that neither model is favored a priori. If $B_{12} > 1$, model M_1 is favored over M_2. The method can be generalized to more than two models. The main difficulty in using the Bayes factor is the evaluation of the integrated likelihood. We used an approximation called the *Bayesian Information Criterion* (BIC) (Schwarz 1978):

$$2 \log p(D|M_k) \approx 2 \log p(D|\hat{\theta}_k, M_k) - \nu_k \log(n) = BIC_k, \tag{4}$$

where ν_k is the number of parameters to be estimated in model M_k, and $\hat{\theta}_k$ is the maximum likelihood estimate for the parameter vector of model M_k, θ_k. For discussion of its use and justification in the context of model-based clustering, see Fraley and Raftery 2000. Intuitively, the first term in Equation (4), which is the maximized mixture likelihood for the model, rewards a model that fits the data well, and the second term discourages overfitting by penalizing models with more free parameters. A large BIC score indicates strong evidence for the corresponding model. Hence, the BIC score can be used to compare models with different covariance matrix parameterizations and different numbers of clusters. Usually, BIC score differences greater than 10 are considered as strong evidence favoring one model over another (Kass and Raftery 1995).

2.3 Prior Work

We are aware of only two published papers attempting model-based formulations of gene expression clustering. Holmes and Bruno (2000) formulate a model that appears to be equivalent to the unconstrained model defined above. Barash and Friedman (2001) define a model similar to the diagonal model above. The main focus of both papers is the incorporation of additional knowledge, specifically transcription factor binding motifs in upstream regions, into the clustering model, and so they do not consider model-based clustering of expression profiles *per se* in the depth or generality that we do. Our results are complementary to those efforts.

3 Data Sets

We used two gene expression data sets for which external evaluation criteria were available, and three sets of synthetic data to test the Gaussian mixture assumption and to compare the performance of different clustering algorithms. We used the term *class* or *component* to refer to a group in the external criterion. The word *cluster* refers to clusters obtained by a clustering algorithm.

3.1 Gene Expression Data Sets

The ovary data: A subset of the ovary data obtained by Schummer, Ng, Bumgarner, Nelson, Schummer, Bednarski, Hassell, Baldwin, Karlan, and Hood (1999) and Schummer (2000) is used. The ovary data set is generated by hybridizing randomly selected cDNA's to membrane arrays. The subset of the ovary data we used contains 235 clones and 24 tissue samples (experiments), some of which are derived from normal tissues, and some from ovarian cancers in various stages of malignancy. The 235 clones were sequenced, and discovered to correspond to 4 different genes. These 4 genes were represented 58, 88, 57, and 32 times on the membrane arrays, respectively. We would hope that clustering algorithms would separate the clones corresponding to these four different genes. Hence, the four genes form the four classes in this data.

The yeast cell cycle data: The yeast cell cycle data (Cho, Campbell, Winzeler, Steinmetz, Conway, Wodicka, Wolfsberg, Gabrielian, Landsman, Lockhart, and Davis 1998) showed the fluctuation of expression levels of approximately 6000 genes over two cell cycles (17 time points). We used two different subsets of this data with independent external criteria. The first subset (the 5-phase criterion) (Cho *et al.* 1998) consists of 384 genes peaking at different time points corresponding to the five phases of cell cycle. Since the 384 genes were identified according to the peak times of genes, we expect clustering results to approximate the five phases. Hence, we used the 384 genes with the 5-phase criterion as one of our data sets. The second subset (the MIPS criterion) consists of 237 genes corresponding to four categories in the MIPS database (Mewes, Heumann, Kaps, Mayer, Pfeiffer, Stocker, and Frishman 1999). The four categories (DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins) were shown to be reflected in clusters from the yeast cell cycle data (Tavazoie, Hughes, Campbell, Cho, and Church 1999).

3.2 Synthetic data sets

Since real expression data sets are expected to be noisy and their clusters may not fully reflect the class information, we complemented our study with synthetic data, for which the classes are known. Modeling gene expression data sets is an ongoing effort by many researchers, and there is no well-established model to represent gene expression data yet. We used the three synthetic data sets proposed by Yeung and Ruzzo (2001). Each of the three synthetic data sets has different properties. By using all three sets of synthetic data, we hope to evaluate the performance of the model-based approach in different situations. The first two synthetic data sets represent attempts to generate replicates of the ovary data set by randomizing different aspects of the original data.

The last synthetic data set is generated by modeling expression data with cyclic behavior. In each of the three synthetic data sets, ten replicates are generated. In each replicate, 235 observations and 24 variables are randomly generated.

Mixture of normal distributions based on the ovary data: Each class in this synthetic data is generated according to a multivariate normal distribution with the sample covariance matrix and the mean vector of the corresponding class in the ovary data. The size of each class in the synthetic data is the same as in the real ovary data. This synthetic data set preserves the mean vector and the covariance matrix between the experiments in each class, but it assumes that the underlying distribution of genes in each class is multivariate normal.

Randomly resampled ovary data: The data for an observation in class c (where $c = 1, \ldots, 4$) under experiment j (where $j = 1, \ldots, 24$) is generated by randomly sampling (with replacement) the expression levels under experiment j in the same class c of the ovary data. The size of each class in this synthetic data set is again the same as in the real ovary data. This data set does not assume any underlying distribution. However, any possible correlation between experiments (for example, the normal tissue samples may be correlated) is not preserved due to the independent random sampling of the expression levels from each experiment. Hence, the resulting sample covariance matrices of this randomly resampled data set tend to be close to diagonal.

Cyclic data: This synthetic data set models cyclic behavior of genes over different time points. The cyclic behavior of genes is modeled by the sine function. Classes are modeled as genes that have similar peak times over the time course. Different classes have different phase shifts and have different sizes. Let x_{ij} be the simulated expression level of gene i under experiment j in this data set with ten classes. Let $x_{ij} = \delta_j + \lambda_j * (\alpha_i + \beta_i \phi(i,j))$, where $\phi(i,j) = \sin(\frac{2\pi j}{8} - w_k + \epsilon)$ (Zhao 2000). The quantity α_i represents the average expression level of gene i, which is chosen according to the standard normal distribution. The quantity β_i is the amplitude control for gene i, which is chosen according to a normal distribution with mean 3 and standard deviation 0.5. The quantity $\phi(i,j)$ models the cyclic behavior. Each cycle is assumed to span 8 time points (experiments). The class number is denoted by k, and the sizes of the different classes are generated according to Zipf's Law (Zipf 1949). Different classes are represented by different phase shifts w_k, which are chosen according to the uniform distribution in the interval $[0, 2\pi]$. The random variable ϵ, which represents the noise of gene synchronization, is generated according to the standard normal distribution. The parameter λ_j is the amplitude control of condition j, and is simulated according to the normal distribution with mean 3 and standard deviation 0.5. The quantity δ_j, which represents an additive experimental error, is generated from the standard normal distribution. Each observation (row) is standardized to have mean 0 and variance 1.

4 Data Transformations and the Gaussian Mixture Assumption

Before applying model-based clustering to gene expression data, we assessed the extent to which the Gaussian mixture assumption holds. Since we do not expect raw expression data to satisfy the Gaussian mixture assumption, we explored the degree of normality of each class after applying different data transformations. In particular, we studied two types of data transformations: the Box-Cox transformations (Box and Cox 1964), and the standardization of each gene (or clone) to have mean 0 and standard deviation 1.

The Box-Cox transformation (Box and Cox 1964) is a parametric family of transformations from y to $y^{(\lambda)}$ with parameter λ:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0. \end{cases} \tag{5}$$

The Box-Cox transformation subsumes many commonly used transformations, including the log transformation which is very popular for microarray data (for example, Speed 2000).

Standardizing each gene (or clone) to have mean 0 and standard deviation 1 is another very popular data transformation step before clustering; see, for example, Tamayo, Slonim, Mesirov, Zhu, Kitareewan, Dmitrovsky, Lander, and Golub (1999) and Tavazoie, Hughes, Campbell, Cho, and Church (1999). Note that this standardization of subtracting the mean and dividing by the standard deviation makes correlation and Euclidean distance equivalent in the transformed data set.

4.1 Methodology for Testing the Gaussian Mixture Assumption

In order to test the Gaussian mixture assumption, the gene expression data sets with external criteria described in Section 3.1 were used. We tested the multivariate normality of *each class* in each data set. There are many different tests for multivariate normality. We used three different approaches: goodness of fit tests based on the empirical distribution function, e.g. Aitchison (1986), skewness and kurtosis tests, e.g. Jobson (1991), and maximum likelihood estimation of the transformation parameters, e.g. Andrews, Gnanadesikan, and Warner (1973).

Aitchison tests: Aitchison (1986) tested three aspects of the data for multivariate normality: the marginal univariate distribution, the bivariate angle distribution and the radius distribution. Suppose a gene expression data set has n genes and p experiments. Since we are interested in clustering the genes, the p experiments are our variables. There are p tests for each of the marginal distributions, $p(p-1)/2$ bivariate angle tests, and one radius test.

Let x_{ij} be the expression level of gene i under experiment j. Suppose the data set has G classes, and class g has n_g genes ($\sum_{g=1}^{G} n_g = n$). Let $\hat{\mu}^{\mathbf{g}} = [\hat{\mu}^{\mathbf{g}}_{\mathbf{j}}]$ and $\hat{\Sigma}^g = [\hat{\sigma}^g_{kj}]$ (where $k, j = 1, \ldots, p$), be the sample mean vector and covariance matrix for class g:

$$\hat{\mu}^g_j = \sum_{i=1}^{n_g} x_{ij}/n_g, \tag{6}$$

$$\hat{\sigma}^g_{kj} = \sum_{i=1}^{n_g} (x_{ik} - \hat{\mu}^g_k)(x_{ij} - \hat{\mu}^g_j)/(n_g - 1). \tag{7}$$

In the **marginal test**, the normality of the marginal distribution of each experiment j is tested. Let $\Phi(\cdot)$ denote the standard normal distribution function, and let $z^g_i = \Phi\{(x_{ij} - \hat{\mu}^g_j)/\sqrt{\hat{\sigma}^g_{jj}}\}$, where $i = 1, \ldots, n_g$. If the x_{ij}'s are normally distributed in class g under experiment j, the sorted values of z^g_i in ascending order should approximate the order statistics of a uniform distribution over the interval (0,1).

Three different statistics computed from the empirical distribution functions (Anderson-Darling, Cramer-won Mises, and Watson) were used to assess departures of the sorted z^g_i values from the order statistics of the uniform distribution. Assuming that z^g_i are the sorted values from class g, The Anderson-Darling statistic is defined by $Q_A = -\{\sum_{i=1}^{n_g}(2i-1)\{\log z^g_i + \log(1 - z^g_{n_g+1-i})\} - n_g\}/n_g$. The Cramer-von Mises statistic is defined by $Q_C = \sum_{i=1}^{n_g}\{z^g_i - (2i-1)/(2n_g)\}^2 + 1/(12n_g)$. The Watson statistics is defined by $Q_W = Q_C - n_g(\bar{z} - \frac{1}{2})^2$ where $\bar{z} = \sum_{i=1}^{n_g} z^g_i/n_g$. Critical values of the empirical distribution function test statistics are given by Aitchison (1986). We used the critical values corresponding to the 1% significance level. For each class, we computed the empirical distribution function test statistics for each of the Anderson-Darling, Cramer-won Mises, and Watson forms using the z^g_i's. If a given test statistic for experiment j is greater than the critical value, we say that the marginal distribution of experiment j shows departure from normality.

In the **bivariate angle test**, the bivariate normality of each pair of experiments (k, j) is evaluated. The idea is that if a pair of variables (u_1, u_2) is circular normal, then the radian angle between the vector from the origin (0,0) to (u_1, u_2) and the u_1-axis is approximately uniform in the interval $[0, 2\pi]$. Since any bivariate normal distribution can be reduced to a circular normal distribution by a suitable transformation, we applied the transformation to each pair of experiments (k, j) and tested the resulting angle for the uniform property. Again, the empirical distribution function test statistics are used to measure the departure from the uniform distribution.

In the **radius test**, the radius of each gene i in class g is defined as $u_i = (\mathbf{x_i} - \hat{\mu}^{\mathbf{g}})^{\mathbf{T}}(\hat{\boldsymbol{\Sigma}}^{\mathbf{g}})^{-\mathbf{1}}(\mathbf{x_i} - \hat{\mu}^{\mathbf{g}})$, where $\mathbf{x_i}$ is the vector of expression levels of gene i under all p experiments. Under the multivariate normal assumption of $\mathbf{x_i}$'s, the radii u_i's are approximately distributed as $\chi^2(p)$. If we define z_i as the sorted values of $F(u_i)$, where F is the distribution function of $\chi^2(p)$, we can again use the empirical distribution function test statistics to measure deviation from the uniform distribution.

Skewness and Kurtosis: Skewness measures the amount of asymmetry in a distribution. For a normal distribution, the skewness is 0. Kurtosis measures the extent to which the data are peaked or flat relative to the normal distribution. For the standard normal distribution, the kurtosis is 3. We computed the skewness and kurtosis of each class g in the data. Let $m_{ir} = (\mathbf{x_i} - \hat{\mu}^{\mathbf{g}})^{\mathbf{T}} (\hat{\mathbf{\Sigma}}^{\mathbf{g}})^{-\mathbf{1}} (\mathbf{x_r} - \hat{\mu}^{\mathbf{g}})$, where $i, r = 1, \ldots, n_g$. Multivariate skewness and kurtosis are defined by $\sum_{i=1}^{n_g} \sum_{r=1}^{n_g} m_{ir}^3 / n_g^2$ and $\sum_{i=1}^{n_g} u_i^2 / n_g$, and null distributions are available for both the multivariate skewness and kurtosis (Mardia 1970). A small p-value suggests the multivariate normal assumption to be questionable.

Maximum likelihood estimation of the transformation parameters: The parameter λ in the Box-Cox transformation in Equation (5) is estimated by maximum likelihood using the observations (Andrews, Gnanadesikan, and Warner 1973). The estimated value of λ suggests both the scale on which the data are closest to normality, and also the extent to which the data on other scales deviate from normality.

4.2 Results of Testing the Gaussian Mixture Assumption

We focused on the popular array data transformations: the logarithmic and square root transformations and the standardization to mean 0 and standard deviation 1. We applied the Aitchison tests and the skewness and kurtosis tests to each class in the transformed ovary data and the transformed yeast cell cycle data. Due to the large number of test statistics from the Aitchison tests $((p + p(p-1)/2 + 1) * 3)$ for each class on any data, only a summary of the Aitchison tests is presented in this technical report. In addition, we found the maximum likelihood estimates of the transformation parameter for each class.

Geometrically, the standardization of subtracting the mean and dividing by the standard deviation of each observation puts the data points on the $(p-2)$-dimensional surface of a $(p-1)$-dimensional sphere. Moreover, the covariance matrices of the standardized data sets are singular. Hence, the skewness and kurtosis tests and the radius test (which involve the inverse of the covariance matrix) are not applicable to the standardized data.

Ovary data: Table 1 shows the results of the Aitchison tests on each of the four classes in the ovary data. In the marginal test, if the test statistics of an experiment j from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level, we adopt the shorthand convention of saying that experiment j *violates* the normality assumption.

The column **m** in Table 1 shows the number of violations from the 24 marginal tests on each class in the ovary data. Similarly, the column **b** in Table 1 shows the number of violations from $\binom{24}{2} = 276$ bivariate angle tests on each class in the ovary data. The column **r** has an entry 1 if the test statistics from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level in the radius test. Otherwise, the column **r** has an entry 0.

Table 1: Results of Aitchison tests on the ovary data. The column m shows the number of violations from the 24 marginal tests on each class. The column b shows the number of violations from 276 bivariate angle tests on each class. The column r has an entry 1 if the test statistics from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level in the radius test, and an entry 0 otherwise.

	class 1			*class 2*			*class 3*			*class 4*		
	m	b	r	m	b	r	m	b	r	m	b	r
raw	0	0	0	5	0	0	18	12	0	4	1	0
log	9	0	0	14	12	0	2	0	0	4	0	0
sqrt	1	0	0	6	0	0	4	0	0	3	0	0
standardized	3	0	NA	7	13	NA	6	0	NA	5	2	NA

The results from Table 1 suggest that the square root transformation is closer to multivariate normal than the log transformation. On the square root transformed data, the marginal test shows that only one experiment (out of 24) deviates from normality in class 1. Similarly, class 2 has 6 experiments, class 3 has 4 experiments and class 4 has 3 experiments that deviate from marginal normality. None of the classes in the square root transformed data shows any deviation in the bivariate angle or radius tests. On the standardized data, the radius tests are not applicable, so the **r** columns for the standardized data are marked “NA” in Table 1.

Table 2 shows the p-values of skewness and kurtosis for each class on the raw, log and square root transformed ovary data. A small p-value indicates deviations from the skewness and kurtosis criteria. From Table 2, class 2 deviates from the skewness and kurtosis criteria in the raw, log and square root transformed data. On the other hand, class 4 does not violate the skewness or kurtosis criteria. Both the square root and log transformations improve skewness in the raw data, but the log transformation makes class 1 skewed. To summarize, the skewness and kurtosis tests show the same overall picture as the Aitchison tests: the square root transformation is relatively close to multivariate normal.

Table 2: p-values of skewness and kurtosis on the ovary data.

		class 1	*class 2*	*class 3*	*class 4*
raw	skewness	0.844	0	0	1
raw	kurtosis	0.999	0.001	0.31	1
log	skewness	0.002	0	0.854	1
log	kurtosis	0.826	0	0.999	1
sqrt	skewness	0.768	0	0.559	1
sqrt	kurtosis	0.999	0.057	0.998	1

Table 3 shows the maximum likelihood estimates of the transformation parameters for each of the four classes in the raw ovary data. $\mathcal{L}_{max}(0.5)$ and $\mathcal{L}_{max}(0)$ are the maximized log-likelihood for the square root and log transformations respectively. From Table 3, the maximum likelihood estimates of the Box-Cox transformation parameter ($\hat{\lambda}$) lie between 0.40 and 0.73 for the four classes in the ovary data. Comparing the maximized log-likelihoods for the square root transformation to those for the log transformation suggests that the square root transformation is more nearly multivariate normal in all four classes.

Table 3: Estimates of the transformation parameter for the ovary data.

class	$\hat{\lambda}$	$\mathcal{L}_{max}(\hat{\lambda})$	$\mathcal{L}_{max}(0.5)$	$\mathcal{L}_{max}(0)$
1	0.728	750	744	678
2	0.658	1195	1188	1060
3	0.405	1221	1219	1179
4	0.590	725	724	689

Yeast cell cycle data with the 5-phase criterion: Table 4 shows the results of the Aitchison tests on the yeast cell cycle data with the 5-phase criterion. The results from Table 4 indicate that the log transformed yeast cell cycle data is more nearly multivariate normal than the square root transformed data. With the log transformation, classes 1, 3, and 4 show no deviation from any of the marginal, bivariate angle and radius tests. The only deviations from normality in this data set are: class 2 shows deviation from the radius test, and one experiment (out of 17) in class 5 shows deviation from marginal normality. The Aitchison tests show that the log transformation greatly enhances normality in all of the 5 classes: the raw data shows significant deviations from the marginal, bivariate angle and radius tests in all of the 5 classes. The standardized yeast cell cycle data is also much more Gaussian than the raw data, but not as much as the log transformed data.

Table 5 portrays a different picture than the Aitchison tests: the raw, square root and log transformed data all show deviations from the skewness and kurtosis criteria. However, the log transformation seems to show relatively less deviation.

Table 4: Results of Aitchison tests on the yeast cell cycle data with the 5-phase criterion. The column headings are as in Table 1.

	class 1			*class 2*			*class 3*			*class 4*			*class 5*		
	m	b	r	m	b	r	m	b	r	m	b	r	m	b	r
raw	17	49	1	17	136	1	17	94	1	17	0	1	17	33	1
log	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
sqrt	8	0	1	17	1	1	15	0	1	0	0	0	7	0	0
standardized	5	0	NA	4	5	NA	1	0	NA	1	0	NA	2	0	NA

Table 5: p-values of skewness and kurtosis on the yeast cell cycle data with the 5-phase criterion.

		class 1	*class 2*	*class 3*	*class 4*	*class 5*
raw	skewness	0	0	0	0	0
raw	kurtosis	0	0	0	0	0
log	skewness	0.051	0	0	0.046	0
log	kurtosis	0.735	0	0	0.678	0.001
sqrt	skewness	0	0	0	0	0
sqrt	kurtosis	0	0	0	0.003	0.001

Table 6 supports the conclusions from the other approaches: the estimated transformation is closer (in terms of differences between estimated Box-Cox power parameters) to the log transformation than to the square root transformation. The estimates $\hat{\lambda}$ are between 0.14 and 0.22 for all 5 classes.

Yeast cell cycle data with the MIPS criterion: In general, the Aitchison tests, the skewness and kurtosis tests, and the maximum likelihood estimation all show similar patterns to the 5-phase criterion: the log transformed data are closer to being Gaussian than the square root transformed data (see Tables 7, 8 and 9). However, class 4 (ribosomal proteins) shows significantly more

Table 6: Estimates of the transformation parameter for the yeast cell cycle data with the 5-phase criterion.

class	$\hat{\lambda}$	$\mathcal{L}_{max}(\hat{\lambda})$	$\mathcal{L}_{max}(0.5)$	$\mathcal{L}_{max}(0)$
1	0.136	-4833	-4910	-4844
2	0.140	-9398	-9591	-9429
3	0.202	-4920	-4975	-4945
4	0.153	-3422	-3468	-3431
5	0.219	-3676	-3713	-3701

deviations from normality with very low p-values for both the skewness and kurtosis tests using the log and square root transformations.

Table 7: Results of Aitchison tests on the yeast cell cycle data with the MIPS criterion. The column headings are as in Table 1.

	class 1			*class 2*			*class 3*			*class 4*		
	m	b	r	m	b	r	m	b	r	m	b	r
raw	17	3	1	17	48	0	17	2	1	9	0	1
log	0	0	0	0	0	0	4	0	1	17	67	1
sqrt	8	0	0	15	0	0	12	0	1	14	1	1
standardized	6	1	NA	2	0	NA	3	0	NA	15	28	NA

Table 8: p-values of skewness and kurtosis on the yeast cell cycle data with the MIPS criterion.

		class 1	*class 2*	*class 3*	*class 4*
raw	skewness	0	0	1	0
raw	kurtosis	0	0.046	1	0
log	skewness	0.136	0.999	1	0
log	kurtosis	0.896	0.999	1	0
sqrt	skewness	0	0.747	1	0
sqrt	kurtosis	0.014	0.996	1	0

Table 9: Estimates of the transformation parameter for the yeast cell cycle data with the MIPS criterion.

$class$	$\hat{\lambda}$	$\mathcal{L}_{max}(\hat{\lambda})$	$\mathcal{L}_{max}(0.5)$	$\mathcal{L}_{max}(0)$
1	0.175	-3448	-3483	-3459
2	0.096	-1912	-1951	-1915
3	0.088	-808	-998	-969
4	0.308	-13188	-13234	-13323

5 Independent Assessment of Clustering Results

The major contribution of this paper is the demonstration of the potential usefulness of the model-based approach, both in terms of the quality of the clustering results and the quality of models selected using the BIC criterion. We compared the performance of the model-based approach to CAST (Ben-Dor and Yakhini 1999), a leading heuristic-based clustering algorithm. Yeung, Haynor, and Ruzzo (2001) compared the performance of several heuristic-based clustering approaches, including several hierarchical clustering algorithms, k-means, and CAST, and concluded that CAST and k-means tend to produce relatively high quality clusters. Since k-means is closely related to maximum likelihood estimation for the equal volume spherical model (EI), we compared the quality of clusters obtained from the model-based approach to that of CAST using correlation as the similarity metric. There is a summary of CAST in Appendix A. In order to assess the clustering results and the number of clusters inferred by the BIC values independently, we used synthetic data sets in which the classes are known and real gene expression sets with external criteria described in Section 3.

5.1 Measure of Agreement

A clustering result can be considered as a partition of objects into groups. Thus, comparing two clustering results is equivalent to assessing the agreement of two partitions. The adjusted Rand index (Hubert and Arabie 1985) assesses the degree of agreement between two partitions. Milligan and Cooper (1986) recommended the adjusted Rand index as the measure of agreement even when comparing partitions with different number of clusters.

Given a set of n objects $S = \{O_1, \ldots, O_n\}$, suppose $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ represent two different partitions of the objects in S such that $\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$ and $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq C$. Suppose that U is our external criterion and V is a clustering result. Let a be the number of pairs of objects that are placed in the same class in U and in the same cluster in V, b be the number of pairs of objects in the same class in U but not in the same cluster in V, c be the number of pairs of objects in the same cluster in V but not in the same class in U, and d be the number of pairs of objects in different classes and different clusters in both partitions. The quantities a and d can be interpreted as agreements, and b and c as disagreements. The Rand index (Rand 1971) is simply $\frac{a+d}{a+b+c+d}$. The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1.

The problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value (say zero). The adjusted Rand index proposed by Hubert and Arabie (1985) assumes the generalized hypergeometric distribution as the model of randomness, *i.e.,* the U and V partitions are picked at random such that the number of objects in the classes and clusters are fixed. Let n_{ij} be the number of objects that are in both class u_i and cluster v_j. Let $n_{i.}$ and $n_{.j}$ be the number of objects in class u_i and cluster v_j respectively. This notation is shown in tabular form in Table 10.

Table 10: Notation for the contingency table for comparing two partitions.

Class or Cluster	v_1	v_2	$\ldots$	v_C	*Sums*
u_1	n_{11}	n_{12}	$\ldots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\ldots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.C}$	$n_{..} = n$

The general form of an index with a constant expected value is $\frac{index - expected\ index}{maximum\ index - expected\ index}$, which is bounded above by 1, and takes the value 0 when the index equals its expected value.

Under the generalized hypergeometric model, it can be shown (Hubert and Arabie 1985) that:

$$E\left[\sum_{i,j}\binom{n_{ij}}{2}\right] = \left[\sum_i \binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right] / \binom{n}{2} \tag{8}$$

The expression $(a + d)$ can be simplified to a linear transformation of $\sum_{i,j}\binom{n_{ij}}{2}$. With simple algebra, the adjusted Rand index (Hubert and Arabie 1985) can be simplified to:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j\binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right] / \binom{n}{2}}. \tag{9}$$

Example 1 in Appendix B illustrates how the adjusted Rand index is computed. The Rand index for comparing the two partitions in Example 1 is 0.711, while the adjusted Rand index is 0.313. The Rand index is much higher than the adjusted Rand index, which is typical. Since the Rand index lies between 0 and 1, the expected value of the Rand index (although not a constant value) must be greater than or equal to 0. On the other hand, the expected value of the adjusted Rand index has value zero and the maximum value of the adjusted Rand index is also 1. Hence, there is a wider range of values that the adjusted Rand index can take on, thus increasing the sensitivity of the index.

6 Results and Discussion

In this section, we show how model-based clustering performed when applied to both gene expression data and to synthetic data based on actual gene expression data. The latter is useful because it has characteristics of gene expression data, but we know what the right answer is, and so we can assess how successful the methods considered are in finding it.

In the model-based approach, parameter estimation becomes difficult when there are too few data points in each component (*i.e.,* too many clusters). Therefore, the BIC values for some of the models are not available when the number of clusters is large. Since CAST is an iterative algorithm with a parameter that indirectly controls the number of clusters produced, the algorithm may not produce a result for every number of clusters. So, in the following result graphs, not all data points are available for CAST.

6.1 Synthetic Data Sets

Mixture of normal distributions based on the ovary data:

Figure 1 shows the average adjusted Rand indices of CAST and four different models using the model-based approach over a range of different numbers of clusters. The average adjusted Rand indices reach the maximum at 4 clusters, with the unconstrained model (VVV) having comparable average adjusted Rand index to CAST. The spherical models (EI and VI) achieve lower quality clustering results than the elliptical models. The diagonal model achieves higher quality clusters than the spherical models on average, but lower than those from the unconstrained model (VVV). Inspection of the covariance matrices of the four classes shows that the covariance matrices are elliptical, and the unconstrained model (VVV) fits the data the best. Hence, our results show the power of the model-based approach when the underlying model is correct.

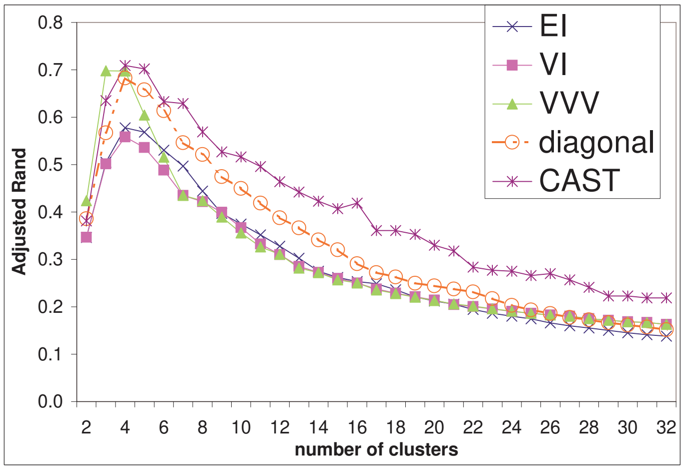

Figure 1: Average adjusted Rand indices for the synthetic data based on the ovary gene expression data: Mixture of 4 multivariate normal distributions

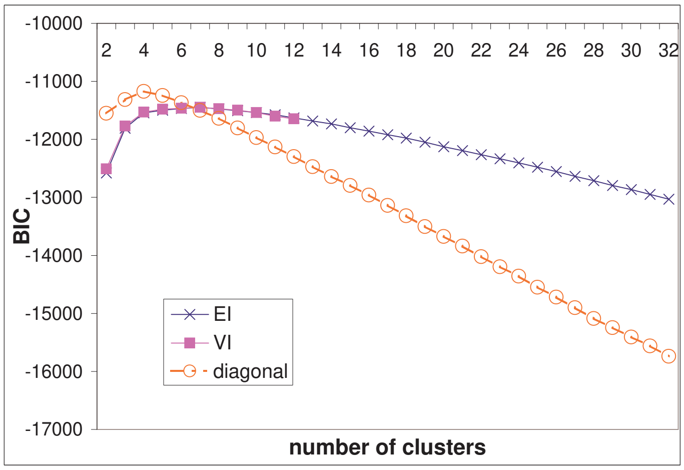

Figure 2: Average BIC values for synthetic data based on the ovary data: Mixture of 4 multivariate normal distributions

Figure 2 shows the average BIC values for three different models using the model-based approach over a range of different numbers of clusters. The BIC values of the unconstrained model (VVV) are not shown because reliable BIC values cannot be computed due to the large number of parameters to be estimated. Nevertheless, with the diagonal model, the maximum average BIC score is reached at 4 clusters, which is the correct number of classes in this data set. In addition, the diagonal model (which achieves higher adjusted Rand indices than the spherical models) also achieves higher BIC values than the spherical models up to 6 clusters. Therefore, the model-based approach favors the diagonal model which produces higher quality clusters.

In order to compare the BIC values of the unconstrained model with the other models, we generated larger synthetic data sets (2350 observations) with the mixture of normal distribution with the mean vectors and covariance matrices of the ovary data. The number of experiments, the number of classes and the relative sizes of the classes remain the same. Our experiments confirmed our hypothesis: with enough data points, the unconstrained model (VVV) produces higher BIC values than the other models, and the maximum BIC score for the unconstrained model is reached at the correct number of classes (4). As for the smaller synthetic data sets, the unconstrained model (VVV) produces higher quality clusters.

Randomly resampled ovary data:

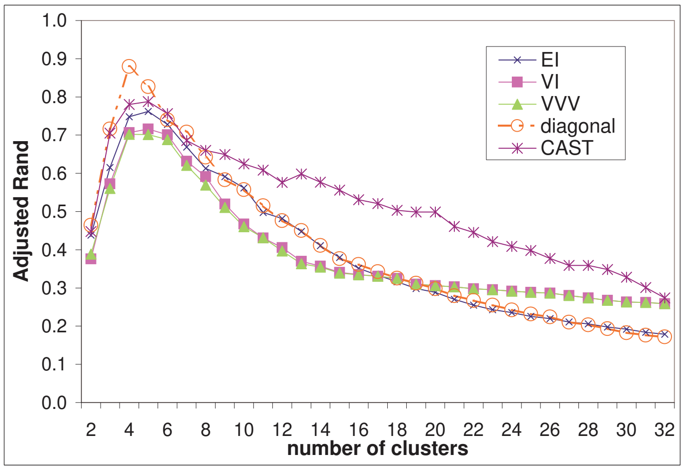

Figure 3: Average adjusted Rand indices for the synthetic randomly resampled ovary data sets.

Figure 3 shows the average adjusted Rand indices for the randomly resampled ovary data sets. The diagonal model achieves clearly superior clustering results compared to other models and to CAST. Figure 4 shows that the BIC analysis selects the diagonal model at the correct number of clusters. Due to the independent sampling of expression levels between experiments, the covariance matrix of each class in this synthetic data set is very close to diagonal. Our results show that the BIC analysis not only selects the right model, but it also determines the correct number of clusters.

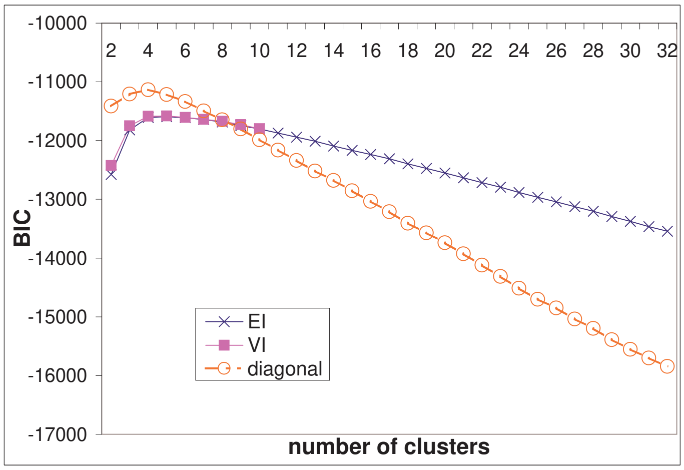

Figure 4: Average BIC values for the synthetic randomly resampled ovary data.

Cyclic data:

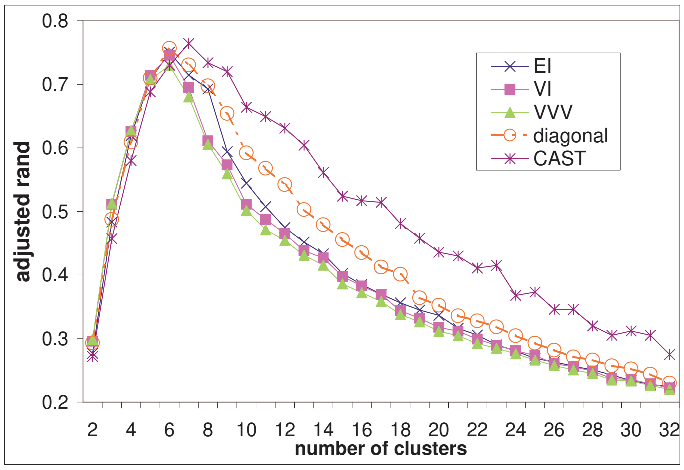

Figure 5: Average adjusted Rand indices for the synthetic cyclic data sets.

Figure 5 shows that the average adjusted Rand indices of CAST and several of the models from the model-based approach are comparable. This synthetic data set contains ten classes. The adjusted Rand indices from CAST are higher than any of the model-based approaches at 10 clusters. In practice, however, one would not know the correct number of clusters, so its performance at the number of clusters that one would select is the most relevant. Furthermore, all of the algorithms show average adjusted Rand indices peaking around 6 or 7 clusters. This set of synthetic data consists of classes with varying sizes, with some very small classes, which can be problematic for most clustering methods including the model-based approach (small clusters make estimation of parameters difficult). In Figure 6, the BIC values of the models also peak around 6 to 7 clusters, with the diagonal model showing higher BIC values (there are too few data points to compute BIC values for the unconstrained model). Our results show that the BIC values select the number of clusters that maximizes the adjusted Rand indices, and the quality of clusters are comparable to CAST at 6 or 7 clusters.

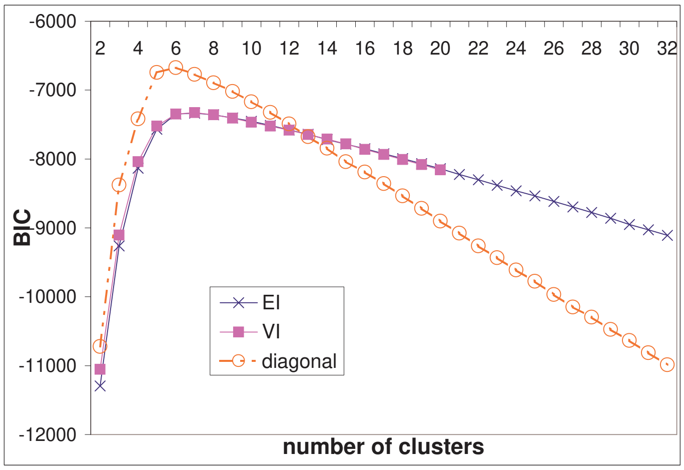

Figure 6: Average BIC values for the synthetic cyclic data sets.

6.2 Gene Expression Data Sets

We compared the clustering results from CAST and the model-based approach on the log transformed, square root transformed, and the standardized ovary data and yeast cell cycle data. A summary of our results is shown in Table 11.

The ovary data:

Since the square root transformation shows the least deviation from the Gaussian mixture assumption, only the results from the square root transformed ovary data are discussed in detail. Figure 7 shows that the spherical models (EI and VI) and the EEE model produce higher quality clusters than CAST and the diagonal and unconstrained models at 4 clusters (the correct number) on the square root transformed ovary data. The BIC curves in Figure 8 show a bend at 4 clusters (which is the number of classes in this data set) for the spherical models, and a maximum at 4 clusters for the diagonal and EEE models. The BIC analysis selects the spherical models at 8 clusters. Even though real expression data may not fully reflect the class structure due to noise, the BIC analysis favors the spherical (EI and VI) and the EEE models over the diagonal models, which is in line with the adjusted Rand indices. Furthermore, closer inspection of the data reveals that the 8 cluster solution selected by BIC analysis is still a meaningful clustering — it differs from the external criterion mainly in that the larger classes have been split into 2 or 3 clusters (which may reflect differences in the constituent cDNAs, for example).

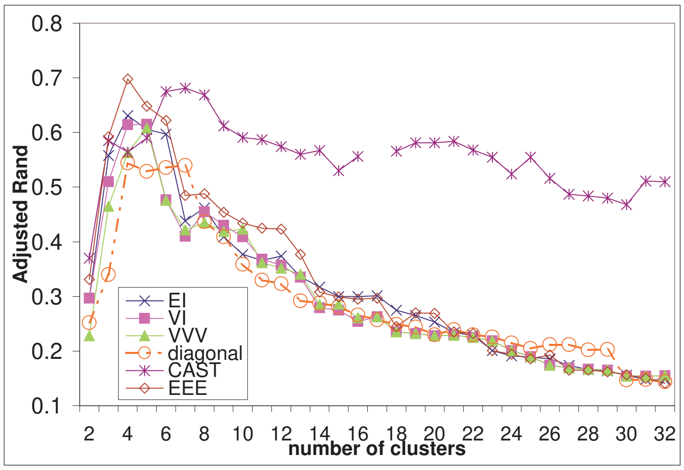

Figure 7: Adjusted Rand indices for the square root transformed ovary data.

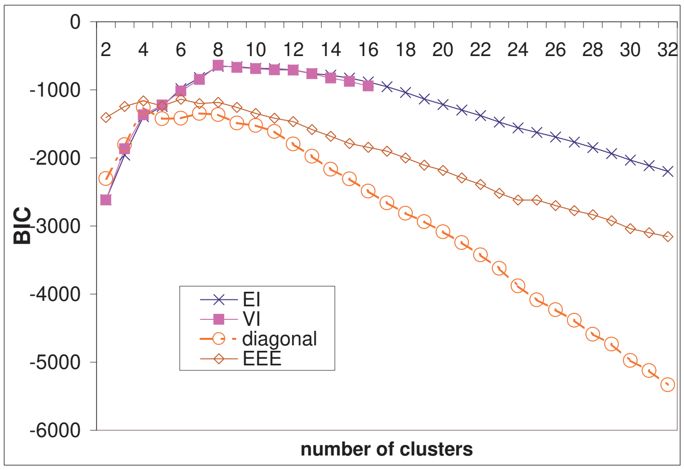

Figure 8: Average BIC values for the square root transformed ovary data.

The results on the log transformed ovary data show that the elliptical models produce clusters with higher adjusted Rand indices than CAST (see Figure 9). The BIC curves on the log transformed ovary data also show a bend at 4 clusters in Figure 10. On the standardized ovary data, the adjusted Rand indices of clusters produced by EEE and EI are comparable to that from CAST (see Figure 11. The BIC curves start to flatten at around 4 clusters on the standardized ovary data, but the maximum occurs at around 7 clusters in Figure 12.

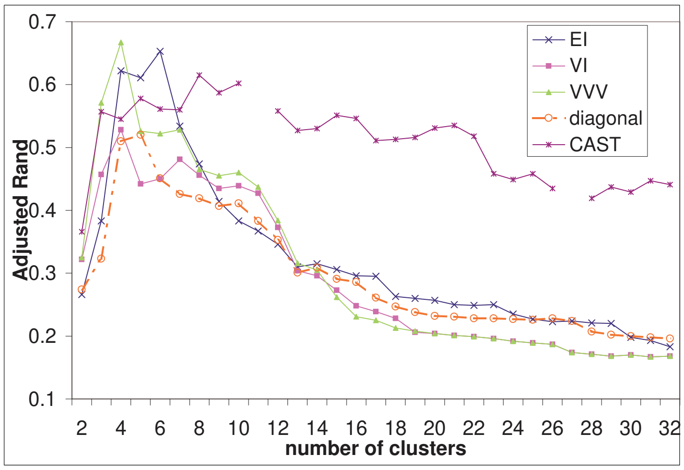

Figure 9: Adjusted Rand indices for the log transformed ovary data.

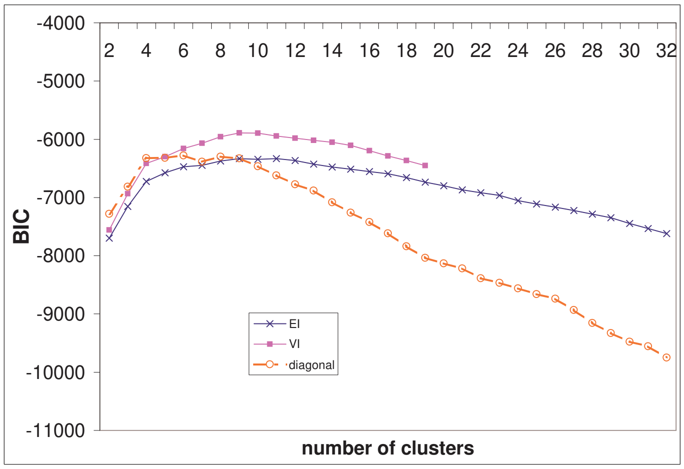

Figure 10: Average BIC values for the log transformed ovary data.

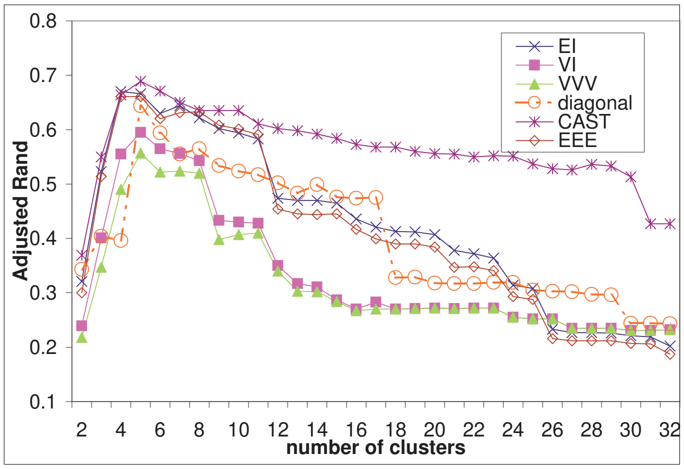

Figure 11: Adjusted Rand indices for the standardized ovary data.

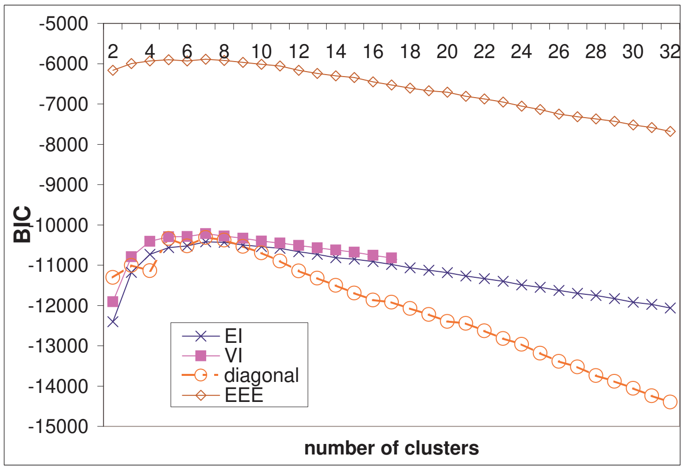

Figure 12: Average BIC values for the standardized ovary data.

Yeast cell cycle data with the 5-phase criterion:

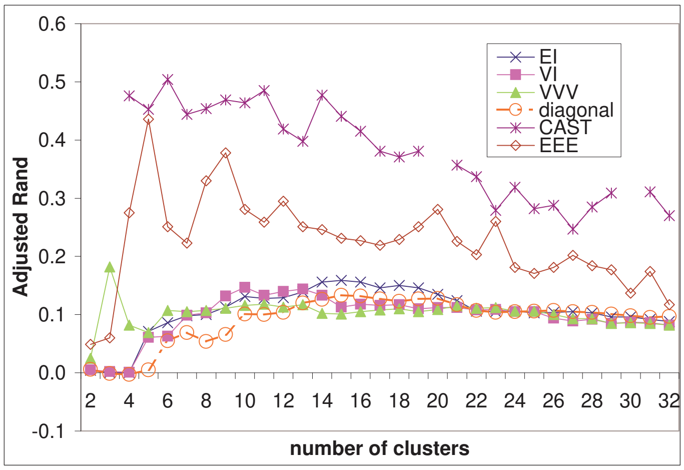

Figure 13: Adjusted Rand indices for the log transformed yeast cell cycle data with the 5-phase criterion.

With the exception of the EEE model, all the other models show considerably lower adjusted Rand indices than those from CAST (see Figure 13) on the log transformed yeast cell cycle data with the 5-phase criterion. Figure 14 shows that the BIC analysis selects the EEE model at 5 clusters (which is the number of classes in this data set). Although the model-based approach on this data set produces lower adjusted Rand indices than CAST, the BIC analysis selects the correct number of clusters and a model with relatively high quality clusters.

The standardized yeast cell cycle data set (Figure 15) shows a very different picture from the log transformed data: the equal volume spherical model (EI) achieves comparable adjusted Rand

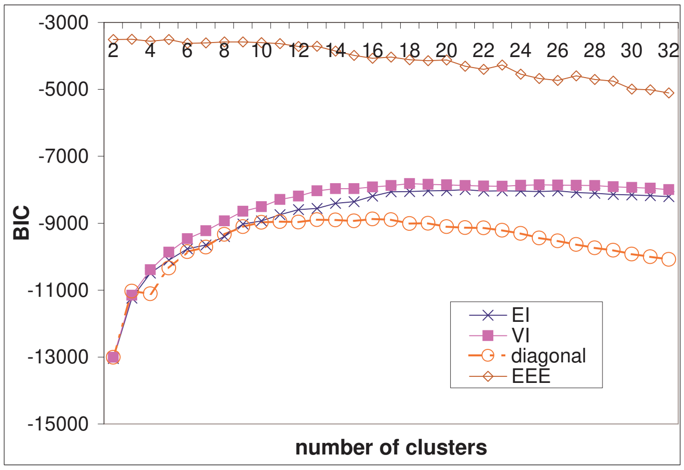

Figure 14: Average BIC values for the log transformed yeast cell cycle data with the 5-phase criterion.

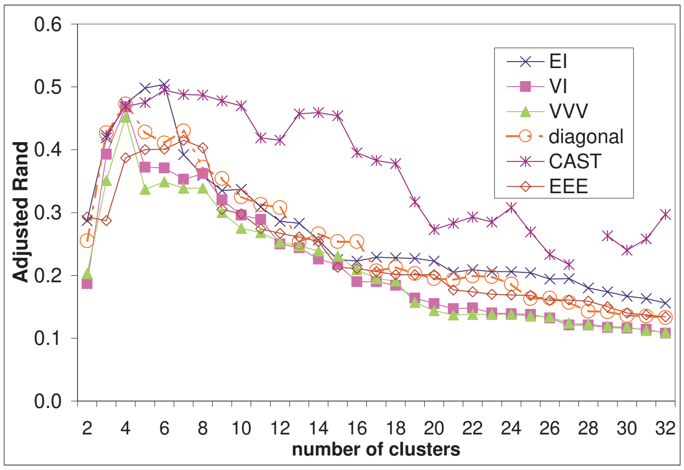

Figure 15: Adjusted Rand indices for the standardized yeast cell cycle data with the 5-phase criterion.

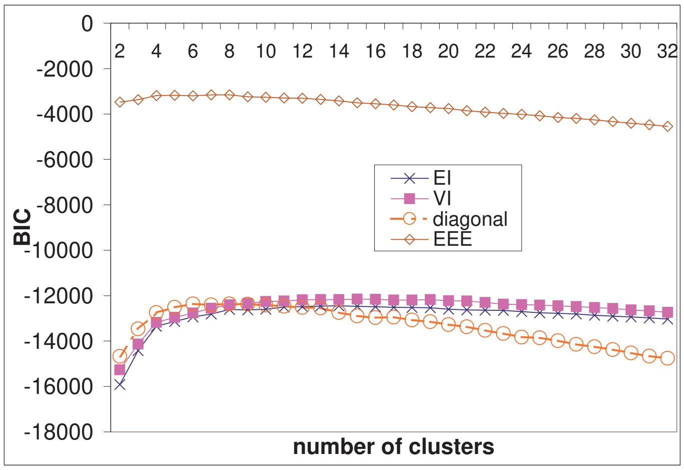

Figure 16: Average BIC values for the standardized yeast cell cycle data with the 5-phase criterion.

indices to CAST at 5 clusters. A careful study of the nature of the data shows that this is no surprise. The yeast cell cycle data set consists of time course data, and so all the 17 experiments are highly correlated (unlike in the ovary data). Figure 17 shows a pairs plot of the first four time points of the log transformed data. Data points from each of the five classes are represented by different symbols in Figure 17. The pairs plots of the remaining 13 time points show a similar pattern. Figure 17 shows that the five classes are not well-separated, and that the data points are scattered along a line. Hence, the model-based approach cannot easily recover the cluster structure. On the contrary, CAST uses correlation coefficients as the similarity measure, and correlation captures the class structure. The five classes have different peak times, and hence have relatively high correlation coefficients within classes compared to correlation between classes. Visualization of the standardized data shows that the data points of the five classes are more spread out and are spherical in shape (see Figure 18). Hence, the model-based approach (in particular, the EI model) is able to capture the class information once the data have been standardized. The BIC analysis (see Figure 16) selects model EEE at 5 clusters.

Yeast cell cycle data with the MIPS criterion:

For the log-transformed yeast cell cycle data with the MIPS criterion, the results are very similar to that with the 5-phase criterion: CAST produces much higher quality clusters than the model-based approach (results not shown). Since this is a different subset of the same data set, the standardization also spreads out the highly correlated data points into spherical clusters, and hence enables the model-based approach to recover the class structure. As for the yeast cell cycle data with the 5-phase criterion, the equal volume spherical (EI) model produces comparable adjusted Rand indices to CAST (see Figure 19) on the standardized data. The BIC curve of model EI shows

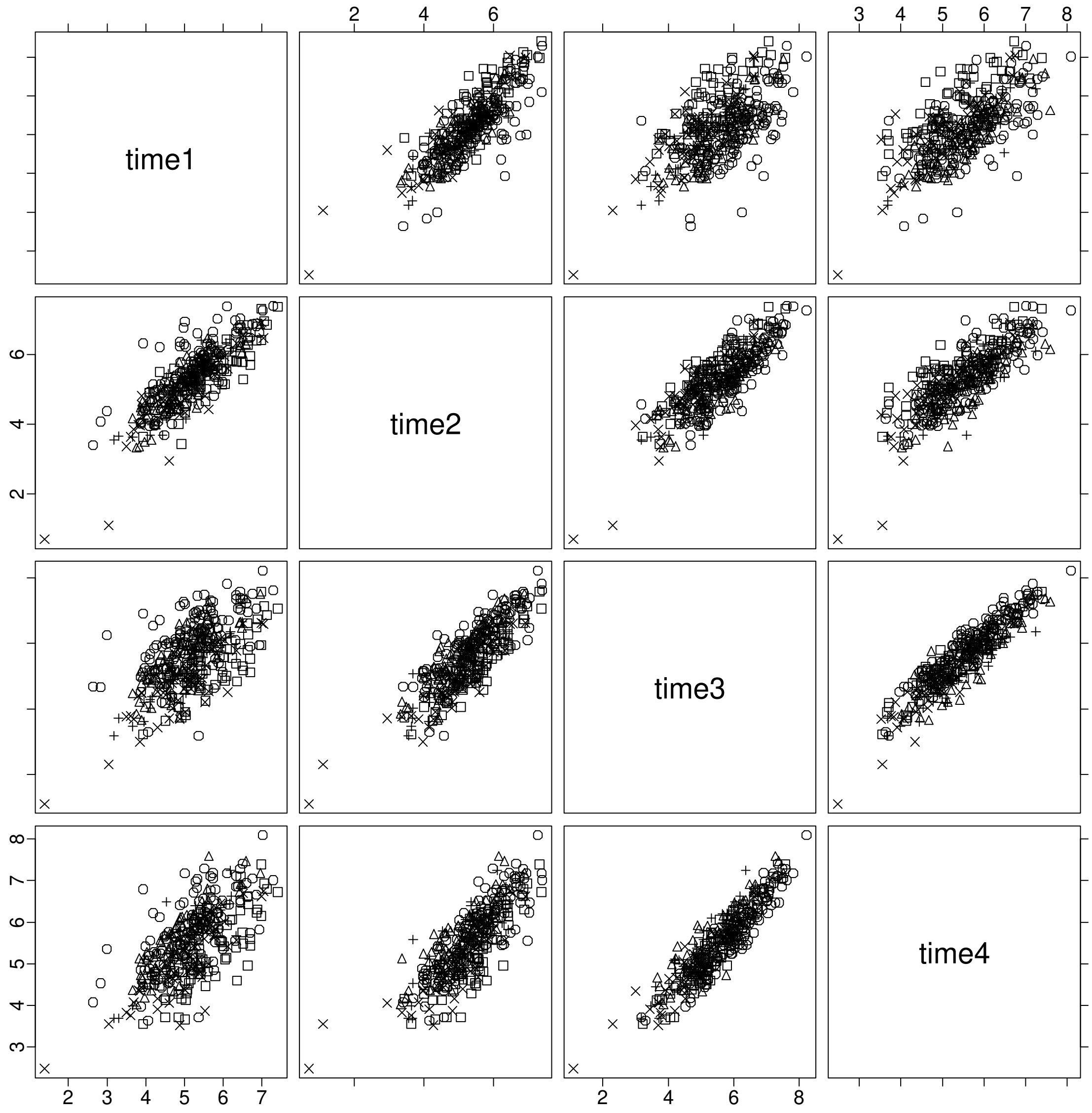

Figure 17: Pairs plot of the log transformed yeast cell cycle data with the 5-phase criterion.

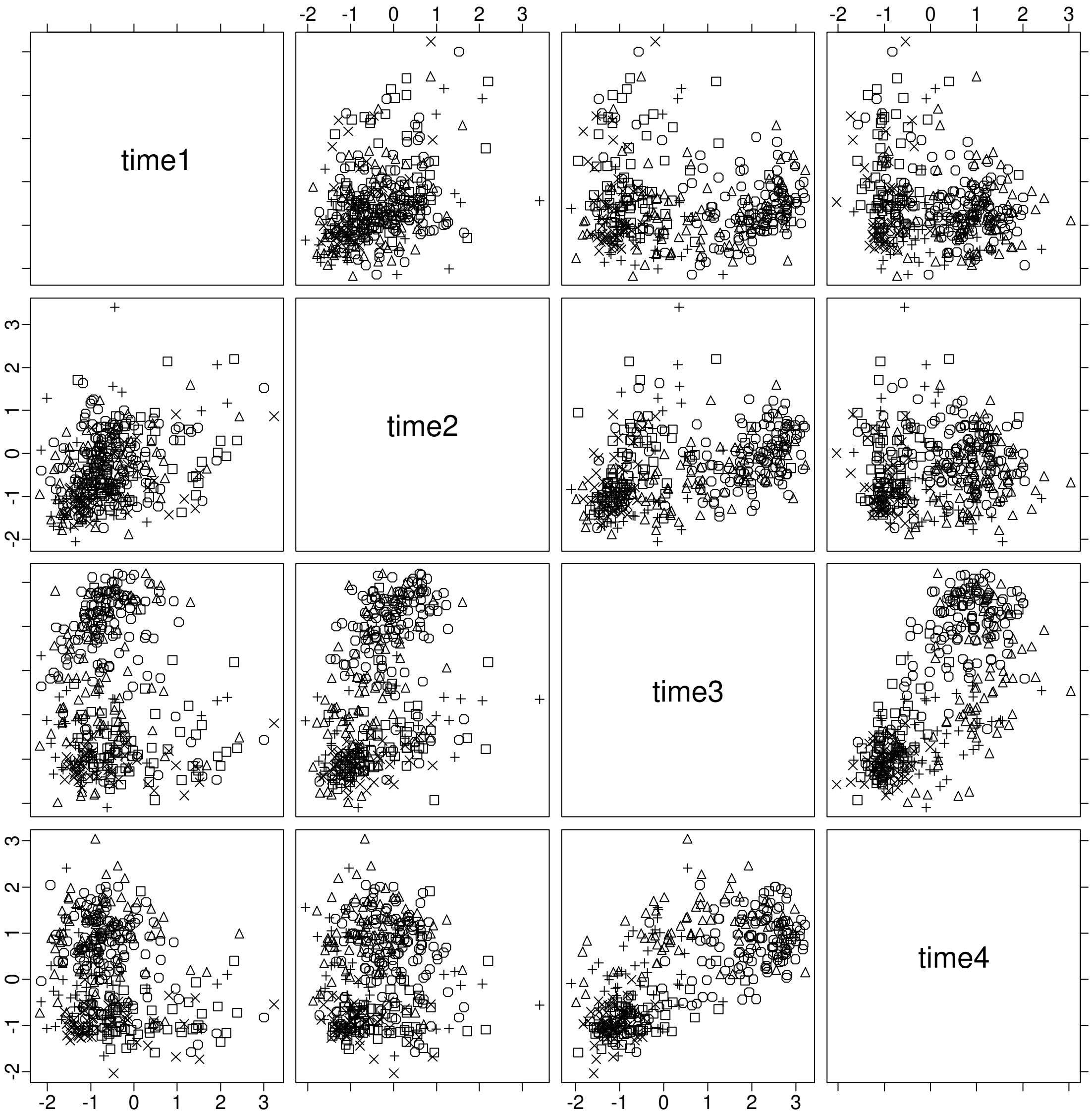

Figure 18: Pairs plot of the standardized yeast cell cycle data with the 5-phase criterion.

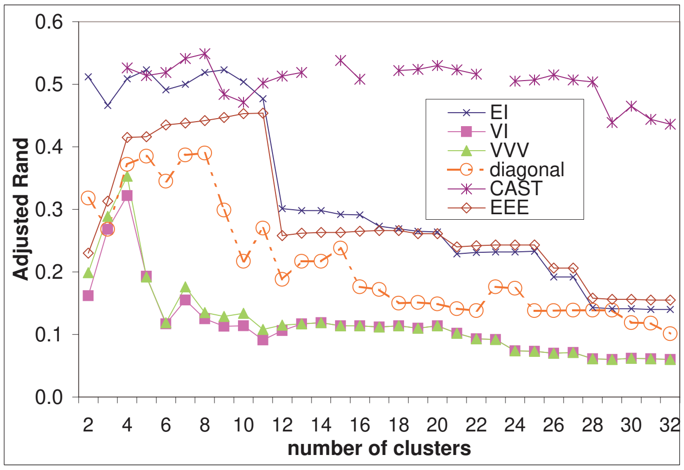

Figure 19: Adjusted Rand indices for the standardized yeast cell cycle data with the MIPS criterion.

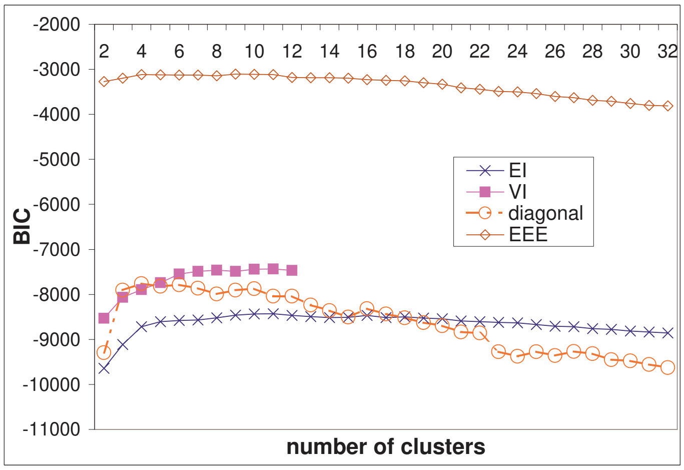

Figure 20: BIC values for the standardized yeast cell cycle data with the MIPS criterion.

Table 11: Summary of results on real expression data.

data set	*# classes*	*Adjusted Rand at # classes*	*model selected by BIC*
log ovary data	4	$EEE > VVV > EI > CAST > VI > diag$	VI at 9 clusters
sqrt ovary data	4	$EEE > EI > VI > CAST > VVV > diag$	VI at 8 clusters
standardized ovary	4	$EI > CAST > EEE > VI > VVV > diag$	EEE at 7 clusters
log 5-phase cell cycle	5	$CAST > EEE \gg EI > VVV > VI > diag$	EEE at 5 clusters
standardized 5-phase	5	$EI > CAST > diag > EEE > VI > VVV$	EEE at 5 clusters
log MIPS cell cycle	4	$CAST > EEE > EI > diag > VI > VVV$	EEE at 4 clusters
standardized MIPS	4	$CAST > EI > EEE > diag > VVV > VI$	EEE at 4 clusters

a bend at 4 clusters (the number of classes in this data set). However, the BIC analysis selects the EEE model at 4 clusters. Note that although the BIC analysis does not select the best model, it does select the correct number of clusters in this data set.

7 Discussion

With our synthetic data sets, the model-based approach had good performance, and also selected the correct model and the right number of clusters using the BIC analysis. For the synthetic normal mixture data sets based on actual gene expression data, the unconstrained model (VVV) produced the highest quality clusters and the BIC analysis chose the right model and the number of clusters when there were enough data points. On the randomly resampled gene expression data sets, the diagonal model produced much higher quality clusters, and the BIC analysis again selected the right model and the correct number of clusters even though the synthetic data sets showed considerable deviation from the Gaussian mixture assumption. On the cyclic data sets (which showed significant deviations from the Gaussian mixture assumption and contained very small classes), we showed that the model-based approach and CAST (a leading heuristic-based approach) produced comparable quality clusters, and the BIC analysis selected the number of clusters that maximized the average adjusted Rand index.

We also showed the practicality of the model-based approach on real gene expression data sets. On the ovary data, the model-based approach achieved slightly better results than CAST, and the BIC analysis gave a reasonable indication of the number of clusters in the transformed data. On two different subsets of the yeast cell cycle data with different external criteria, the equal volume spherical model (EI) and EEE model produced comparable results to CAST on the standardized data. The BIC values from the EEE model were maximized at the correct number of clusters.

In addition, we showed that data transformations can increase the extent to which gene expression data sets are consistent with normality, and that models have varying performance on data sets that are transformed differently. Although real expression data sets do not perfectly satisfy the Gaussian mixture assumption even after various data transformations, the model-based approach nevertheless produces slightly higher quality clusters, and provides an indication of the numbers of clusters. It is interesting to note that simple models, like the equal volume spherical model (EI) and the elliptical EEE model, produced relatively high quality clusters on all of our transformed data sets. The EEE model even determined the right number of clusters on two different subsets from the yeast cell cycle data set with different external criteria. On the ovary data set, the BIC values overestimated the number of clusters and did not select the model with the highest adjusted Rand indices. However, inspection of the clusters showed that the clustering result selected by the BIC analysis is still meaningful.

In our study, we found different data transformations helpful in clustering different types of expression data. For gene expression data sets with different tissue types (for example, the ovary data), we found the log or the square root transformation to work best. For time course data sets or other highly correlated data sets, we recommend standardizing the data set and applying the equal volume spherical model (EI) to cluster the data sets. When k-means or the equal volume spherical model (EI) is used, we believe standardizing the data set before the clustering step will generally be useful.

Overall, our results suggest the potential usefulness of model-based clustering even with existing implementations, which are not tailored for gene expression data sets. We believe that custom refinements to the model-based approach could be of value for gene expression analysis. There are many directions for such refinements. One direction is to design models that incorporate specific information about the experiments. For example, for expression data sets with different tissue types (like the ovary data), the covariances among tissue samples of the same type are expected to be higher than those between tissue samples of different types. Hence, a block matrix parameterization of the covariance matrix would be a reasonable assumption. Another advantage of customized parameterizations of the covariance matrices is that the number of parameters to be estimated could be greatly reduced. Another crucial direction of future research is to incorporate missing data and outliers in the model. We believe that the overestimation of the number of clusters in the ovary data may be due to noise or outliers. In this paper, we used subsets of data without any missing values. With the underlying probability framework, we expect the ability to model outliers and missing values explicitly to be another potential advantage of the model-based approach over the heuristic clustering methods.

In terms of data transformations, more types of data transformations can be explored. In particular, the data transformation may depend on the technology with which the array data is produced. For example, the ovary data set used in this study is produced by the membrane arrays, while the yeast cell cycle data set is produced by the Affymetrix arrays. Different array technologies may produce data with different statistical properties.

Acknowledgements

We would like to thank Michèl Schummer from the Institute of Systems Biology for providing us with the ovary data set, and for helpful discussions about its interpretation. We would also like to thank Trey Ideker, Roger Ngouenet, Saurabh Sinha, Jeremy Tantrum, and Vesteinn Thorsson, for helpful discussions. The research of Yeung and Ruzzo was partially supported by NSF grant DBI-9974498. The research of Raftery and Fraley was supported by Office of Naval Research grants N00014-96-1-0192 and N00014-96-1-0330. These analyses were carried out using the model-based clustering software MCLUST available at www.stat.washington.edu/fraley/mclust/soft.shtml.

APPENDIX

A Details of the CAST algorithm

The Cluster Affinity Search Technique (CAST) is an algorithm proposed by Ben-Dor and Yakhini (1999) to cluster gene expression data. The input to the algorithm includes the pairwise similarities of the genes, and a cutoff parameter t (which is a real number between 0 and 1). The clusters are constructed one at a time. The current cluster under construction is called C_{open}. The *affinity* of a gene g, $a(g)$, is defined to be the sum of similarity values between g and all the genes in C_{open}. A gene g is said to have high affinity if $a(g) \geq t|C_{open}|$. Otherwise, g is said to have low affinity. Note that the affinity of a gene depends on the genes that are already in C_{open}. The algorithm alternates between adding high affinity genes to C_{open}, and removing low affinity genes from C_{open}. C_{open} is *closed* when no more genes can be added to or removed from it. Once a cluster is closed, it is not considered any more by the algorithm. The algorithm iterates until all the genes have been assigned to clusters and the current C_{open} is closed.

When a new cluster C_{open} is started, the initial affinity of all genes are 0 since C_{open} is empty. One additional heuristic that Ben-Dor and Yakhini (1999) implemented in their software BIOCLUST is to choose a gene with the maximum number of neighbors to start a new cluster. Another heuristic is that after the CAST algorithm converges, there is an additional iterative step, in which all clusters are considered at the same time, and genes are moved to the cluster with the highest average similarity.

B Example illustrating the adjusted Rand index

The following example illustrates how the adjusted Rand index (discussed in Section 5.1) is computed. Example 1 is a contingency table in the same form as in Table 10.

Example 1

Class or Cluster	v_1	v_2	v_3	*Sums*
u_1	1	1	0	2
u_2	1	2	1	4
u_3	0	0	4	4
Sums	2	3	5	$n = 10$

a is defined as the number of pairs of objects in the same class in U and same cluster in V, hence a can be written as $\sum_{i,j} \binom{n_{ij}}{2}$. In Example 1, $a = \binom{2}{2} + \binom{4}{2} = 7$. b is defined as the number of pairs of objects in the same class in U but not in the same cluster in V. In terms of the notation in Table 1, b can be written as $\sum_i \binom{n_{i.}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$ In Example 1, $b = \binom{2}{2} + \binom{4}{2} + \binom{4}{2} - 7 = 6$. Similarly, c is defined as the number of pairs of objects in the same cluster in V but not in the same class in U, so c can be written as $\sum_j \binom{n_{.j}}{2} - \sum_{i,j} \binom{n_{ij}}{2} = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} - 7 = 7$. d is defined as the number of pairs of objects that are not in the same class in U and not in the same cluster in V. Since $a + b + c + d = \binom{n}{2}$, $d = \binom{10}{2} - 7 - 6 - 7 = 25$. The Rand index for comparing the two partitions in Example 1 is $\frac{7+25}{45} = 0.711$, while the adjusted Rand index is $\frac{7-14*13/45}{(14+13)/2-14*13/45} = 0.313$ (see Section 5.1 for the definitions of the Rand and adjusted Rand indices). The Rand index is much higher than the adjusted Rand index.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.

Andrews, D. F., R. Gnanadesikan, and J. L. Warner (1973). Methods for assessing multivariate normality. In P. R. Krishnaiah (Ed.), *Multivariate analysis III*, pp. 95–116. New York: Academic Press.

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

Barash, Y. and N. Friedman (2001, April). Context-specific bayesian clustering for gene expression data. In T. Lengauer et al. (Eds.), *RECOMB 2001, Proceedings of the Fifth Annual International Conference on Computational Biology*, Montreal, Canada, pp. 12–21.

Ben-Dor, A. and Z. Yakhini (1999, March). Clustering gene expression patterns. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France, pp. 33–42.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *J. R. Statist. Soc. B 26*, 211–252.

Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, M. Ares, and D. Haussler (2000). Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Science USA 97*(1), 262–267.

Celeux, G. and G. Govaert (1992). A classification em algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis 14*, 315–332.

Celeux, G. and G. Govaert (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation 47*, 127–146.

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *The Journal of the Pattern Recognition Society 28*, 781–793.

Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis (1998, July). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell 2*, 65–73.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998, December). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA 95*, 14863–14868.

Fraley, C. and A. E. Raftery (1998). How many clusters? which clustering method? - answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (1999). Mclust: Software for model-based cluster analysis. *Journal of Classification 16*, 297–306. Available at www.stat.washington.edu/tech.reports/tr342.ps.

Fraley, C. and A. E. Raftery (2000, October). Model-based clustering, discriminant analysis, and density estimation. Technical Report 380, University of Washington, Department of Statistics. Available at www.stat.washington.edu/tech.reports.

Hartuv, E., A. Schmitt, J. Lange, S. Meirer-Ewert, H. Lehrach, and R. Shamir (1999, March). An algorithm for clustering cdnas for gene expression analysis. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France, pp. 188–197.

Holmes, I. and W. J. Bruno (2000, August). 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*, La Jolla, CA, pp. 202–210. AAAI Press.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* (2), 193–218.

Jobson, J. D. (1991). *Applied multivariate data analysis*. New York: Springer-Verlag.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Lander, E. S. (1999). Array of hope. *Nature Genetics 21*, 3–4.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika 57*, 519–530.

McLachlan, G. J. and K. E. Basford (1988). *Mixture models: inference and applications to clustering*. Marcel Dekker New York.

Mewes, H. W., K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman (1999). Mips: a database for protein sequences and complete genomes. *Nucleic Acids Research 27*, 44–48.

Milligan, G. W. and M. C. Cooper (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research 21*, 441–458.

Murua, A., J. Tantrum, W. Stuetzle, and S. Sieberts (2001). Model based document classification and clustering. Manuscript in preparation.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 846–850.

Schummer, M. (2000). Manuscript in preparation.

Schummer, M., W. V. Ng, R. E. Bumgarner, P. S. Nelson, B. Schummer, D. W. Bednarski, L. Hassell, R. L. Baldwin, B. Y. Karlan, and L. Hood (1999). Compartive hybridization of an array of 21500 ovarian cdnas for the discovery of genes overexpressed in ovarian carcinomas. *An International Journal on Genes and Genomes 238*, 375–385.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Speed, T. P. (2000). Speed group microarray page: Hints and prejudices. http://stat-www.berkeley.edu/users/terry/zarray/Html/hintsindex.html.

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Science USA 96*, 2907–2912.

Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church (1999). Systematic determination of genetic network architecture. *Nature Genetics 22*, 281–285.

Tibshirani, R., G. Walther, and T. Hastie (2000). Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Dept. of Statistics, Stanford University.

Yeung, K. Y., D. R. Haynor, and W. L. Ruzzo (2001). Validating clustering for gene expression data. *Bioinformatics 17*(4), 309–318. Available at http://www.cs.washington.edu/homes/kayee/cluster/.

Yeung, K. Y. and W. L. Ruzzo (2001). An empirical study on principal component analysis for clustering gene expression data. *To appear in Bioinformatics*. Available at http://www.cs.washington.edu/homes/kayee/pca/.

Zhao, L. P. (2000). Personal communications.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.