BIOINFORMATICS

Vol. 17 no. 10 2001
Pages 977–987

Model-based clustering and data transformations for gene expression data

K. Y. Yeung [1,*], C. Fraley [2], A. Murua [3], A. E. Raftery [2] and W. L. Ruzzo [1]

[1]Computer Science and Engineering, Box 352350, [2]Statistics, Box 354322, University of Washington, Seattle, WA 98195, USA and [3]Insightful Corporation, 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA

Received on April 20, 2001; accepted on July 6, 2001

ABSTRACT

Motivation: Clustering is a useful exploratory technique for the analysis of gene expression data. Many different heuristic clustering algorithms have been proposed in this context. Clustering algorithms based on probability models offer a principled alternative to heuristic algorithms. In particular, model-based clustering assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. 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. Gaussian mixture models have been shown to be a powerful tool for clustering in many applications.

Results: 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 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 explored the validity of the Gaussian mixture assumption on different transformations of real data. 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.

Availability: MCLUST is available at http://www.stat.washington.edu/fraley/mclust. The software for the diagonal model is under development.

Contact: kayee@cs.washington.edu

Supplementary information: http://www.cs.washington.edu/homes/kayee/model

*To whom all correspondence should be addressed.

1 INTRODUCTION AND MOTIVATION

DNA microarrays offer the promise of studying the variation of many genes simultaneously. Researchers have generated large amounts of gene expression data, but there is a great need to develop analytical methodology to analyze and to exploit this information (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 *et al.*, 1998), self-organizing maps (Tamayo *et al.*, 1999), k-means (Tavazoie *et al.*, 1999), graph-theoretic approaches (for example, Ben-Dor and Yakhini, 1999; Hartuv *et al.*, 1999), and support vector machines (Brown *et al.*, 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 *et al.* (1998) and Tamayo *et al.* (1999) used visual display to determine the number of clusters. Yeung *et al.* (2001b) suggested clustering the data set leaving out one experiment at a time and then comparing the performance of different clustering algorithms using the left-out experiment. The gap statistic (Tibshirani *et al.*, 2000) estimates the number of clusters by comparing within-cluster dispersion to that of a reference null distribution. However, in the absence of a well-grounded statistical model, it seems difficult to define what is meant by a 'good' clustering algorithm or the 'right' number of clusters.

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 probabil-

ity 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). This provides a great advantage over heuristic clustering algorithms, for 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 explored the extent to which different transformations of gene expression data sets satisfy the normality assumption. Due to space limitations, data transformations and normality tests are summarized in Section 5.2, and described in Yeung *et al.* (2001a) and our supplementary web site.

In Section 5, we show that the existing model-based clustering implementations produce higher quality clustering results than a leading heuristic approach when the data 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.

Our contributions include demonstrations of the potential usefulness of the model-based approach by testing the Gaussian mixture assumption for different transformations of expression data, applying existing model-based clustering implementations to both real expression data and synthetic data sets, and comparing the performance of the model-based approach to a leading heuristic-based algorithm.

2 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

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

where f_k and θ_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 $\left(\tau_k \geqslant 0 \text{ and } \sum_{k=1}^{G} \tau_k = 1\right)$.

In the Gaussian mixture model, each component k is modeled by the multivariate normal distribution with parameters μ_k (mean vector) and Σ_k (covariance matrix):

$$f_k(\mathbf{y}_i | \mu_k, \Sigma_k) = \frac{\exp\{-\frac{1}{2}(\mathbf{y}_i - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{y}_i - \mu_k)\}}{\sqrt{\det(2\pi \Sigma_k)}}. \qquad (2)$$

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 exploiting the representation of the covariance matrix in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^{\mathrm{T}}, \qquad (3)$$

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 its shape, and λ_k determines its volume.

Allowing some but not all of the parameters in equation (3) to vary results in a set of models within this general framework that is sufficiently flexible to accommodate data with widely varying characteristics. In this paper, we consider five such models, outlined below. Constraining $D_k A_k D_k^{\mathrm{T}}$ to be the identity matrix I corresponds to Gaussian mixtures in which each component is spherically symmetric. The *equal volume spherical* model (denoted 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, requiring relatively more data points in each component. There are a range of elliptical models with other constraints and fewer parameters. For example, with the parameterization $\Sigma_k = \lambda D A D^{\mathrm{T}}$, each component is elliptical, but all have equal volume, shape and orientation (denoted 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 EI, VI, EEE and

VVV models as implemented in MCLUST (Fraley and Raftery, 1999), and the diagonal model as implemented by Murua *et al.* (2001).

In both the MCLUST implementation and the diagonal model implementation, the desired number of clusters G is specified, and then the model parameters (τ_k, μ_k and Σ_k appropriately constrained, for $1 \leqslant k \leqslant G$) are estimated by the EM algorithm. In the EM algorithm, the 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 conditional probabilities from EM provide a 'soft clustering,' since a data point may have nonzero probability of belonging to several clusters at once, but we do not pursue any analysis that uses this information here.) In the clustering context, the EM algorithm for mixture models 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 very closely related to model-based clustering using the EI model, 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, given k-means' interpretation as a parsimonious model of simple independent Gaussians, which is adequate to describe data arising in many contexts. However, there are circumstances in which the model underlying k-means may *not* be appropriate. For example, the VI model would make more sense 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. We have also 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 distinguish between these scenarios (and others). However, there is a tradeoff in that the more general models require more parameters to be estimated. In the worst case—that of allowing the orientation to vary between clusters—there are $\theta(p^2)$ parameters to be estimated per cluster, where p is the number of variables in the data. Another key advantage of model-based clustering is that there is a principled, data-driven way to approach the model selection problem, which 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 a model selection problem. 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 $p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)\,\mathrm{d}\theta_k$ 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):

$$2\log p(D|M_k) \approx 2\log p(D|\widehat{\theta_k}, M_k) - \nu_k \log(n) = \mathrm{BIC}_k \tag{4}$$

where ν_k is the number of parameters to be estimated in model M_k, and $\widehat{\theta_k}$ is the maximum likelihood estimate for parameter θ_k. 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. (The formal derivation of the BIC approximation does not rely on this intuition.) 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 incorporation of additional knowledge, specifically transcription factor binding motifs in upstream regions, into the clustering model, and so 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 compare the performance of different clustering algorithms. We use 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

3.1.1 Ovary data. We used a subset of the ovary data obtained by Schummer *et al.* (1999); Schummer (2000). The ovary data set was generated by hybridization to a randomly selected cDNA (clone) library arrayed on nylon membranes. 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 four different genes. These four genes were represented 58, 88, 57 and 32 times on the membrane arrays, respectively. Ideally, clustering algorithms should separate the clones corresponding to these four different genes. Hence, the four genes form the four classes in this data.

3.1.2 Yeast cell cycle data. The yeast cell cycle data (Cho *et al.*, 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 five-phase criterion) consists of 384 genes whose expression levels peak at different time points corresponding to the five phases of cell cycle (Cho *et al.*, 1998). We expect clustering results to approximate this five class partition. Hence, we used the 384 genes with the five-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 *et al.*, 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 *et al.*, 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 yet. We used the three synthetic data sets proposed in 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 scenarios. The first two synthetic data sets replicate different aspects of the original ovary data set. The last synthetic data set models expression data with cyclic behavior.

3.2.1 Mixture of normal distributions based on the ovary data. Each class in this synthetic data was generated according to a multivariate normal distribution with the sample covariance matrix and the mean vector of the corresponding class in the standardized ovary data (each gene in a standardized data set has mean 0 and standard deviation 1, see Section 5.2 for more details). A total of 2350 observations were generated for each data replicate. The size of each class in the synthetic data is ten times that of the corresponding class in the 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 expression levels in each class is multivariate normal.

3.2.2 Randomly resampled ovary data. In contrast to the previous synthetic data set, this one preserves the marginal empirical distributions of the real ovary data, but not its covariance structure. Specifically, the value for an observation in class c (where $c = 1, \ldots, 4$) under experiment j (where $j = 1, \ldots, 24$) was generated by randomly sampling (with replacement) the expression levels under the same experiment j in the same class c from the standardized ovary data. The size of each class in this synthetic data set is the same as in the real ovary data. Due to the independent random sampling of the expression levels from each experiment, any possible correlation between experiments (for example, the normal tissue samples may be correlated) is lost. Hence, the resulting sample covariance matrix of each class from this synthetic data set is close to diagonal.

3.2.3 Cyclic data. This synthetic data set models sinusoidal cyclic behavior of genes over time. Classes are modeled as genes that have similar peak

times (phase shifts) over the time course. Let x_{ij} be the simulated expression level of gene i under experiment j, where $i = 1 \ldots 235$ and $j = 1 \ldots 24$. 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). α_i represents the average expression level of gene i, which is chosen according to the standard normal distribution. 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. $\phi(i, j)$ models the cyclic behavior. Each cycle is assumed to span eight time points (experiments). There are a total of ten classes, and k is the class number. 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 (gene) is standardized to have mean 0 and variance 1.

4 INDEPENDENT ASSESSMENT OF CLUSTERS

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 compare the performance of the model-based approach to CAST (Ben-Dor and Yakhini, 1999), a leading heuristic-based clustering algorithm. Yeung *et al.* (2001b) compared the performance of many 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 the EM algorithm for the EI model, we compared the quality of clusters obtained from the model-based approach to that of CAST using correlation as the similarity metric. A summary of CAST can be found in Yeung *et al.* (2001a). In order to assess the clustering results and the number of clusters inferred by the BIC scores independently, we used synthetic data sets in which the classes are known as well as real gene expression sets with external criteria described in Section 3.

4.1 Measure of agreement

A clustering result can be considered as a partition of objects into groups. Thus, comparing a clustering result to the external criterion 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. Based on an extensive empirical study, Milligan and Cooper (1986) recommended the adjusted Rand index as the measure of agreement even when comparing partitions with different numbers of clusters. In this paper, we used the adjusted Rand index to assess the clustering results by comparing to the corresponding external criterion.

The *Rand index* (Rand, 1971) is defined as the fraction of agreement, i.e. the number of pairs of objects that are either in the same groups in both partitions or in different groups in both partitions, divided by the total number of pairs of objects. The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1. The *adjusted* Rand index (Hubert and Arabie, 1985) adjusts the score so that its expected value in the case of random partitions is 0. A high adjusted Rand index indicates a high level of agreement between the two partitions. Please refer to Yeung *et al.* (2001a) for a detailed description of the adjusted Rand index.

5 RESULTS AND DISCUSSION

In this section, we show how model-based clustering performed when applied to both synthetic and real gene expression data. Due to space limitations, only selected results are shown. Please refer to our supplementary web site or Yeung *et al.* (2001a) for more results. In the model-based approach, parameter estimation becomes difficult when there are too few data points in each cluster. As a result, the BIC scores of some of the models are not available when the number of clusters is large. For example, with the VVV model, there are $p + p(p + 1)/2$ parameters to be estimated per cluster, where p is the dimension of the data. With the ovary data, we have $p = 24$, so that the number of parameters to be estimated for the VVV model is 324, which is greater than the number of observations (235) in the data set. Even for the mixture of normal distributions based on the ovary data with 2350 observations, when the number of clusters is greater than 7, the number of parameters to be estimated for the VVV model would exceed the number of data points (2350). 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.

5.1 Synthetic data sets

In this subsection, we present results from our synthetic data sets. In each case, the results presented are the average values over ten replicates.

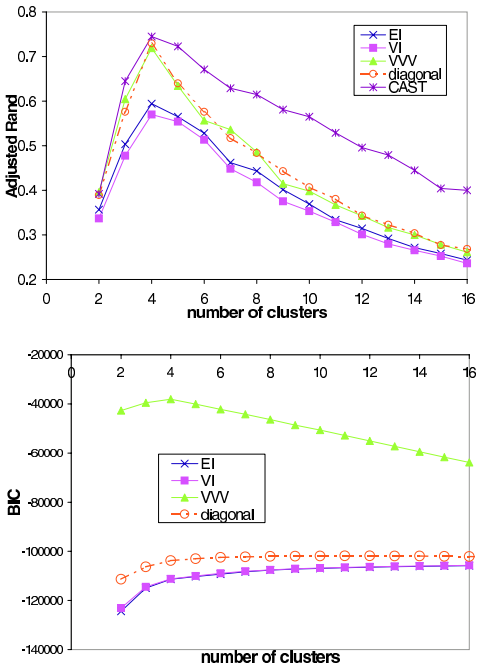

Fig. 1. Average adjusted Rand indices (top) and average BIC scores (bottom) for the mixture of normal synthetic data.

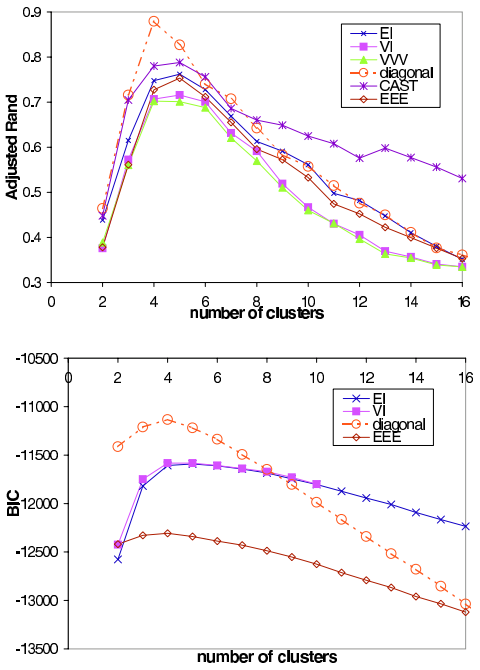

Fig. 2. Average adjusted Rand indices (top) and average BIC scores (bottom) for the randomly resampled ovary data.

5.1.1 Mixture of normal distributions based on the ovary data. Figure 1 (top) 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 four clusters, with the VVV model, the diagonal model and CAST having comparable average adjusted Rand indices. The spherical models (EI and VI) achieve lower quality clustering results than the elliptical models. Inspection of the covariance matrices of the four classes shows that the covariance matrices are elliptical, and the VVV model fits the data the best.

Figure 1 (bottom) shows the average BIC scores of four different models using the model-based approach over a range of different numbers of clusters. The maximum average BIC score is achieved by the VVV model at four clusters, which is the number of classes in this data set. Moreover, the diagonal model produces higher BIC scores than the spherical models, which is in line with the results from the adjusted Rand index. Therefore, the BIC analysis selects the right model and the correct number of clusters on this synthetic data set.

5.1.2 Randomly resampled ovary data. Figure 2 (top) shows the average adjusted Rand indices for the randomly resampled ovary data. The diagonal model achieves clearly superior clustering results compared to other models and CAST. Figure 2 (bottom) shows that the BIC analysis selects the diagonal model at the correct number of clusters (4). 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 also determines the correct number of clusters.

5.1.3 Cyclic data. Figure 3 (top) 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 ten 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 six or seven 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 3 (bottom), the BIC scores of the models also peak

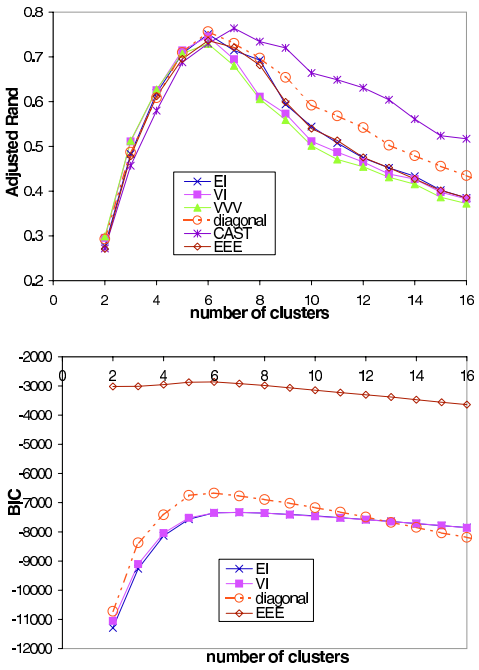

Fig. 3. Average adjusted Rand indices (top) and average BIC scores (bottom) for the cyclic data.

around 6–7 clusters, with the EEE model showing higher BIC scores (there are too few data points to compute BIC scores for the VVV model). Our results show that the BIC scores select the number of clusters that maximizes the adjusted Rand indices, and the quality of clusters from the model-based approach is comparable to CAST at 6 or 7 clusters.

5.2 Data transformations and the Gaussian mixture assumption

The model-based clustering methods we have tried all assume that the data come from a Gaussian mixture. In practice, the methods are reasonably robust to deviations from this model (see the results above, for example) but work best when the data are, or can be transformed to be, nearly Gaussian. Consequently, we tested the validity of the Gaussian mixture assumption for both the ovary data (radiolabeled cDNAs on nylon membranes) and the yeast data (Affymetrix oligo arrays). We considered both the raw expression values and the data values after applying each of three commonly used transformations: logarithm, square root, and standardization (wherein the raw expression levels for each gene are transformed by subtracting their mean and dividing by their standard deviation).

We used a variety of tests suggested by Aitchison (1986) to assess multivariate normality in each class. As the simplest example, we tested whether the marginal distributions for any of the experiments deviated from (univariate) normality. More elaborate tests looked at the angular distribution of points in each pair of experiments and the radial distribution of points from all experiments. We also applied the skewness and kurtosis tests (Jobson, 1991) in which the symmetry and flatness of the distribution are compared to the normal distribution. From our results, the square root transformed ovary data showed relatively less deviation from the Gaussian mixture assumption. For example, although one of the four classes failed to satisfy the skewness and kurtosis tests (under all of the transformations, in fact), and a few of the experiments failed the marginal tests for each class, the square root transformed data satisfied all the remaining tests at the 1% significance level. For the yeast cell cycle data with both criteria, the log transform satisfied the Gaussian mixture assumption relatively well. For all three data sets, the raw values fit the Gaussian model poorly, hence suitable data transformation is an important step in analysis of such data. In the interest of space, detailed results are omitted; please refer to Yeung *et al.* (2001a) or our supplementary web site.

5.3 Gene expression data sets

5.3.1 Ovary data. Figure 4 (top) shows that the spherical models (EI and VI) and the EEE model produce higher quality clusters than CAST and the diagonal and VVV models at four clusters (which is the correct number of classes) on the square root transformed ovary data. However, the rate of decline of the adjusted Rand index from CAST is less steep than that from the model-based approach so that the adjusted Rand index from CAST is higher than that from the model-based approach when the number of clusters is large. In Figure 4 (bottom), the EEE model has its first local maximum BIC score at four clusters (the correct number of classes), the diagonal model has its global maximum BIC score at four clusters, and the BIC curves of the spherical models (EI and VI) show a bend at four clusters. However, the spherical models (EI and VI) at eight clusters achieve the highest BIC scores. Even though real expression data may not fully reflect the class structure due to noise, the BIC analysis favors the EEE model over the diagonal model, 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 two or three clusters (which may reflect differences in the constituent cDNAs, for example).

The results on the log transformed ovary data show that the elliptical models produce clusters with higher adjusted Rand indices than CAST. The BIC curves on the log transformed ovary data also show a bend at four clusters

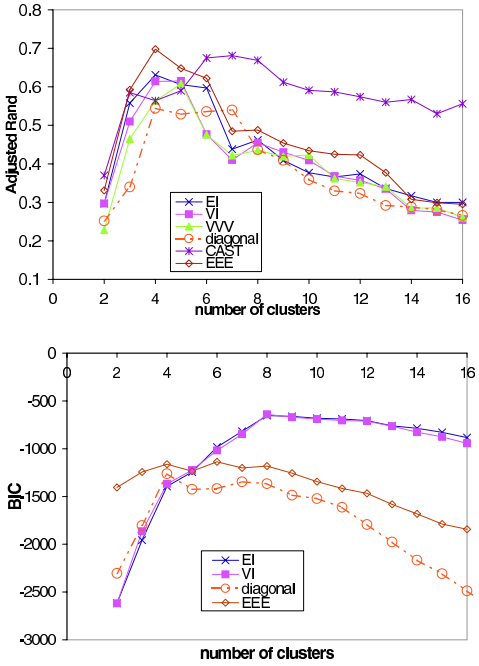

Fig. 4. Adjusted Rand indices (top) and BIC scores (bottom) for the square root transformed ovary data.

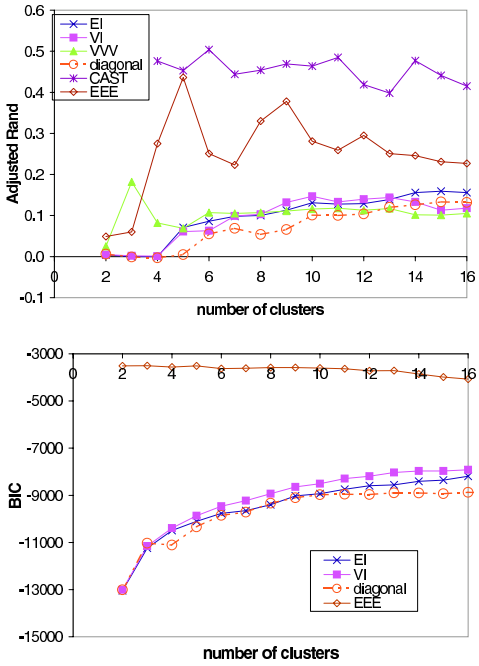

Fig. 5. Adjusted Rand indices (top) and BIC scores (bottom) for the log transformed yeast cell cycle data with the five-phase criterion.

(figures not shown). On the standardized ovary data, the adjusted Rand indices of clusters produced by EEE and EI are comparable to that from CAST. The BIC curves start to flatten at around four clusters on the standardized ovary data, but the maximum occurs at around seven clusters.

5.3.2 Yeast cell cycle data with the five-phase criterion. With the exception of the EEE model, all the other models show considerably lower adjusted Rand indices than those from CAST (Figure 5 top) on the log transformed yeast cell cycle data with the five-phase criterion. Figure 5 (bottom) shows that the BIC analysis selects the EEE model at five clusters, which is the number of classes in this data. 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 6) shows a very different picture from the log transformed data: the EI model achieves comparable adjusted Rand indices to CAST at five 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 17 experiments are highly correlated (unlike the ovary data). Visualization of the log transformed data shows that the five classes are not well-separated, and the data points are scattered along a line. Hence, the model-based approach cannot easily recover the class structure. The classes in this data set are based on peak times of the five phases of cell cycle, and so the classes capture the 'general patterns' across the experiments and not the absolute expression levels of the genes. Standardization captures this information better than log transformation. In contrast, CAST uses correlation coefficients as the similarity measure, and correlation captures the 'general patterns' across experiments even when the data set is log-transformed (without standardization). Visualization of the standardized data shows that the data points of each of the five classes are more spread out and are spherical in shape. Hence, the EI model captures the class information on the standardized data. The BIC analysis (figure not shown) selects model EEE at five clusters.

In addition, we experimented with another data transformation that captures the 'general patterns' across the experiments. Specifically, we took the logarithm of the ratio of the expression level of a gene to the total expression level of the gene over all experiments, i.e. $\log\left(x_{ij}/\sum_{k=1}^{17} x_{ik}\right)$, where x_{ij} is the expression level of gene i under experiment j. The results of this transformation are similar to those from standardization (figure not shown): the model-based approach achieves

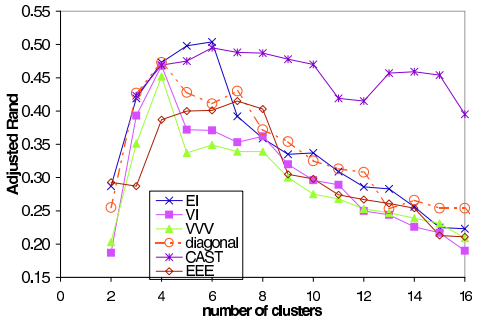

Fig. 6. Adjusted Rand indices for the standardized yeast cell cycle data with the five-phase criterion.

comparable adjusted Rand indices to CAST. Hence, if the goal of clustering is to capture the 'general patterns' across experiments and not the absolute expression levels, the data set should be appropriately transformed to reflect this objective before applying model-based clustering.

5.3.3 Yeast cell cycle data with the MIPS criterion. For the yeast cell cycle data with the MIPS criterion, the results are very similar to that with the five-phase criterion: CAST produces much higher quality clusters than the model-based approach for log-transformed data (figure not shown). However, the model-based approach works well on standardized data—standardization again captures the class structure, and hence enables the model-based approach to recover the class structure. As with the yeast cell cycle data with the five-phase criterion, the EI model produces comparable adjusted Rand indices to CAST (Figure 7 top) on the standardized data. The BIC curve of model EI shows a bend at four clusters, which is the number of classes in this data (Figure 7, bottom). However, the BIC analysis selects the EEE model at four clusters. Note that although the BIC analysis does not select the best model, it does select the second-best model and the correct number of clusters in this data set. Furthermore, careful inspection shows that the clustering result selected by the BIC analysis still captures most of the class information.

6 CONCLUSIONS

6.1 Summary

With our synthetic data sets, the model-based approach not only showed superior performance, but also selected the correct model and the right number of clusters using the BIC analysis. On the mixture of normal distribution synthetic data sets, the VVV model produced the highest quality clusters and the BIC analysis chose the right model and the number of clusters. On the randomly resampled

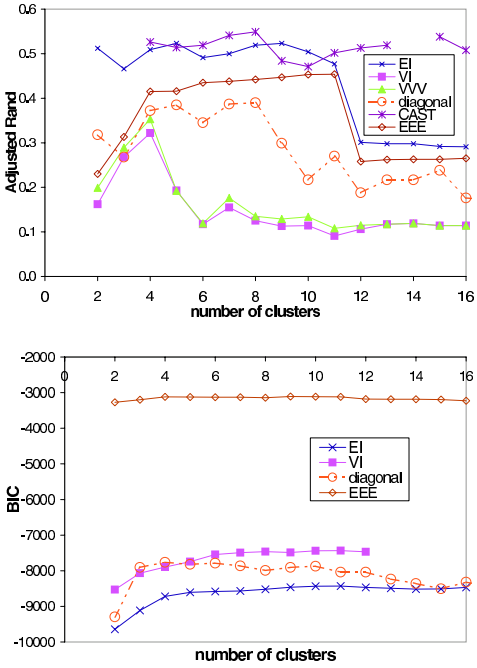

Fig. 7. Adjusted Rand indices (top) and BIC scores (bottom) for the standardized yeast cell cycle data with the MIPS criterion.

synthetic data sets with close to diagonal covariance matrices, 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 EI model and EEE model produced comparable results to CAST on the standardized data. The BIC scores from the EEE model were maximized at the correct number of clusters. The results are summarized in Table 1.

Table 1. Summary of results on real expression data

Data	No. of classes	Transformation	Maximum adjusted Rand at no. of classes	BIC analysis	
				Model selected	Notes
Ovary	4	log	EEE	VI at nine clusters	EI, VI, diagonal: bend at four clusters
Ovary	4	sqrt	EEE	VI at eight clusters	EEE: local max at four clusters
Ovary	4	standardized	EI	EEE at seven clusters	EI, VI, EEE: bend at four clusters
Five-phase cell cycle	5	log	EEE, CAST	EEE at five clusters	
Five-phase cell cycle	5	standardized	EI	EEE at five clusters	
MIPS cell cycle	4	log	CAST	EEE at four clusters	
MIPS cell cycle	4	standardized	CAST, EI	EEE at four clusters	

The method giving the highest adjusted Rand index at the number of classes is shown in the fourth column. When the adjusted Rand indices from two methods are approximately the same, two methods are shown.

6.2 Conclusions

We showed that data transformations can greatly enhance normality in expression data sets, and 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 suggests the numbers of clusters. It is interesting to note that simple models, like the EI model 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 scores 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 nevertheless meaningful.

In our study, we showed that data sets should be appropriately transformed to reflect the goal of clustering. In particular, if the goal is to capture the general patterns across experiments without considering the absolute expression levels, data transformations such as standardization are helpful.

6.3 Future work

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 would be of great 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 on 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.

ACKNOWLEDGEMENTS

We would like to thank Michèl Schummer from the Institute of Systems Biology for the ovary data set. We would also like to thank Trey Ideker, Roger Ngouenet, Saurabh Sinha, Jeremy Tantrum, and Vesteinn Thorsson. Yeung and Ruzzo are partially supported by NSF grant DBI-9974498. The research by Raftery and Fraley was supported by Office of Naval Research grants N00014-96-1-0192 and N00014-96-1-0330.

REFERENCES

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

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

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

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

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

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

Celeux,G. and Govaert,G. (1993) Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Stat. Comput. Simul.*, **47**, 127–146.

Celeux,G. and Govaert,G. (1995) Gaussian parsimonious clustering models. *J. Pattern Recognition Soc.*, **28**, 781–793.

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

Dasgupta,A. and Raftery,A.E. (1998) Detecting features in spatial point processes with clutter via model-based clustering. *J. Am. Stat. Assoc.*, **93**, 294–302.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Fraley,C. and Raftery,A.E. (1998) How many clusters? Which clustering method?—answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

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

Hartuv,E., Schmitt,A., Lange,J., Meirer-Ewert,S., Lehrach,H. and Shamir,R. (1999) 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 Bruno,W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. In Altman,R. *et al.* (ed.), *Proceedings Eighth Annual International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, La Jolla, CA, pp. 202–210.

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

Jobson,J.D. (1991) *Applied Multivariate Data Analysis*. Springer, New York.

Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

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

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

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

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

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

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

Schummer,M. (2000) Manuscript in preparation.

Schummer,M., Ng,W.V., Bumgarner,R.E., Nelson,P.S., Schummer,B., Bednarski,D.W., Hassell,L., Baldwin,R.L., Karlan,B.Y. and Hood,L. (1999) Compartive hybridization of an array of 21 500 ovarian cdnas for the discovery of genes overexpressed in ovarian carcinomas. *Int. J. Genes Genomes*, **238**, 375–385.

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

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

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

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

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, to appear. Available at http://www.cs.washington.edu/homes/kayee/pca/.

Yeung,K.Y., Fraley,C., Murua,A., Raftery,A.E. and Ruzzo,W.L. (2001a) Model-based clustering and data transformations for gene expression data. *Technical Report UW-CSE-01-04-02*, Department of Computer Science and Engineering, University of Washington. Available at our supplementary web site.

Yeung,K.Y., Haynor,D.R. and Ruzzo,W.L. (2001b) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

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

Zipf,G.K. (1949) *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, MA.