Principal Component Analysis for clustering gene expression data

Yeung, K. Y.*
Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195, USA

Ruzzo, W. L.
Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195, USA

July 11, 2001

running head: an empirical study on principal component analysis for clustering expression data

key words: gene expression analysis, clustering, principal component analysis, dimension reduction

To appear in *Bioinformatics*

*To whom correspondence should be addressed

Abstract

Motivation: *There is a great need to develop analytical methodology to analyze and to exploit the information contained in gene expression data. Because of the large number of genes and the complexity of biological networks, clustering is a useful exploratory technique for analysis of gene expression data. Other classical techniques, such as principal component analysis (PCA), have also been applied to analyze gene expression data. Using different data analysis techniques and different clustering algorithms to analyze the same data set can lead to very different conclusions. Our goal is to study the effectiveness of principal components (PC's) in capturing cluster structure. Specifically, using both real and synthetic gene expression data sets, we compared the quality of clusters obtained from the original data to the quality of clusters obtained after projecting onto subsets of the principal component axes.*

Results: *Our empirical study showed that clustering with the PC's instead of the original variables does not necessarily improve, and often degrade, cluster quality. In particular, the first few PC's (which contain most of the variation in the data) do not necessarily capture most of the cluster structure. We also showed that clustering with PC's has different impact on different algorithms and different similarity metrics. Overall, we would not recommend PCA before clustering except in special circumstances.*

Availability: *The software is under development.*

Contact: *kayee@cs.washington.edu*

Supplementary information:

http://www.cs.washington.edu/homes/kayee/pca

1 Introduction and Motivation

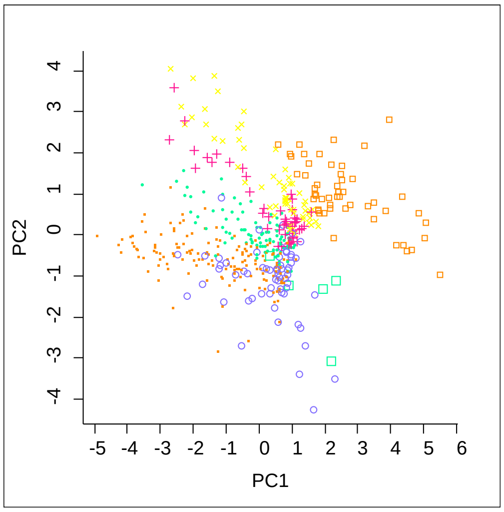
DNA microarrays offer the first great hope to study variations of many genes simultaneously (Lander, 1999). Large amounts of gene expression data have been generated by researchers. There is a great need to develop analytical methodology to analyze and to exploit the information contained in gene expression data (Lander, 1999). Because of the large number of genes and the complexity of biological networks, clustering is a useful exploratory technique for analysis of gene expression data. Many clustering algorithms have been proposed for gene expression data. For example, (Eisen *et al.*, 1998) applied a variant of the hierarchical average-link clustering algorithm to identify groups of co-regulated yeast genes. (Ben-Dor and Yakhini, 1999) reported success with their CAST algorithm.

Other techniques, such as principal component analysis (PCA), have also been proposed to analyze gene expression data. PCA (Jolliffe, 1986) is a classical technique to reduce the dimensionality of the data set by transforming to a new set of variables (the principal components) to summarize the features of the data. Principal components (PC's) are uncorrelated and ordered such that the kth PC has the kth largest variance among all PC's. The kth PC can be interpreted as the direction that maximizes the variation of the projections of the data points such that it is orthogonal to the first $k-1$ PC's. The traditional approach is to use the first few PC's in data analysis since they capture most of the variation in the original data set. In contrast, the last few PC's are often assumed to capture only the residual "noise" in the data. PCA is closely related to a mathematical technique called singular value decomposition (SVD). In fact, PCA is equivalent to applying SVD on the covariance matrix of the data. Recently, there has been a lot of interest on applying SVD to gene expression data, for example, (Holter *et al.*, 2000) and (Alter *et al.*, 2000).

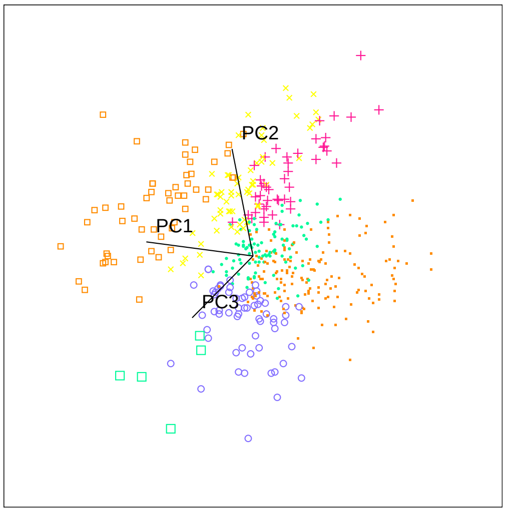
Using different data analysis techniques and different clustering algorithms to analyze the same data set can lead to very different conclusions. For example, (Chu *et al.*, 1998) identified seven clusters in a subset of the sporulation data set using a variant of the hierarchical clustering algorithm of (Eisen *et al.*, 1998). However, (Raychaudhuri *et al.*, 2000) reported that these seven clusters are very poorly separated when the data is visualized in the space of the first two PC's, even though they account for over 85% of the variation on the data.

PCA and clustering: In the clustering literature, PCA is sometimes applied to reduce the dimensionality of the data set prior to clustering. The hope for using PCA prior to cluster analysis is that PC's may "extract" the cluster structure in the data set. Since PC's are uncorrelated and ordered, the first few PC's, which contain most of the variations in the data, are usually used in cluster analysis, for example, (Jolliffe *et al.*, 1980). There are some common rules of thumb to choose how many of the first PC's to retain, but most of these rules are informal and ad-hoc (Jolliffe, 1986). On the other hand, there is a theoretical result showing that the first few PC's may not contain cluster information: assuming that the data is a mixture of two multivariate normal distributions with different means but with an identical within-cluster covariance matrix, (Chang, 1983) showed that the first few PC's may contain less cluster structure information than other PC's. He also generated an artificial example in which there are two clusters, and if the data points are visualized in two dimensions, the two clusters are only well-separated in the subspace of the first and last PC's.

A motivating example: A subset of the sporulation data (477 genes) were classified into 7 temporal patterns (Chu *et al.*, 1998). Figure 1(a) is a visualization of this data in the space of the first 2 PC's, which contains 85.9% of the variation in the data. Each of the seven patterns is represented by a different color or different shape. The seven patterns overlap around the origin in Figure 1(a). However, if we view the same subset of data points in the space of the first 3 PC's



(a) In the subspace of the first 2 PC's



(b)In the subspace of the first 3 PC's

Figure 1: Visualization of a subset of the sporulation data

(containing 93.2% of the variation in the data) in Figure 1(b), the seven patterns are much more separated. This example shows that a small variation (7.4%) in the data helps to distinguish the patterns, and different numbers and different sets of PC's have varying degree of effectiveness in capturing cluster structure. Therefore, there is a great need to investigate the effectiveness of PCA as a preprocessing step to cluster analysis on gene expression data before one can identify clusters in the space of the PC's. This paper is an attempt of such an empirical study.

2 Our approach

Our goal is to empirically investigate the effectiveness of clustering gene expression data using PC's instead of the original variables. In this paper, genes are clustered, hence the experimental conditions are the variables. Our methodology is to run a clustering algorithm on a given data set, and then apply the same algorithm to the data after projecting it into the subspaces defined by different sets of PC's. The effectiveness of clustering with the original data and with different sets of PC's is determined by assessing the quality of clusters, which is measured by comparing the clustering results to an objective external criterion of the data. In our experiments, we assume the number of clusters is known and clustering results with the correct number of clusters are produced. Both real gene expression data sets with external criteria and synthetic data sets are used in this empirical study.

2.1 Agreement between two partitions

In order to compare clustering results against external criteria, a measure of agreement is needed. The adjusted Rand index (Hubert and Arabie, 1985) assesses the degree of agreement between two partitions of the same set of objects. Based on an extensive empirical comparison of several such measures, (Milligan and Cooper, 1986) recommended the adjusted Rand index as the measure of agreement even when comparing partitions having different numbers 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 $1 \leq j \neq j' \leq C$. In our case, one of the partitions is the external criterion and one is a clustering result. Let a be the number of pairs of objects that are placed in the same element in partition U and in the same element in partition V, and d be the number of pairs of objects in different elements in partitions U and V. The Rand index (Rand, 1971) is simply the fraction of agreement, *i.e.,* $(a+d)/\binom{n}{2}$. The Rand index lies between 0 and 1. When the two partitions are identical, the Rand index is 1. A problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value. The adjusted Rand index (Hubert and Arabie, 1985) corrects for this by assuming the general form $\frac{index - expected\ index}{maximum\ index - expected\ index}$. Its maximum value is 1 and its expected value in the case of random clusters is 0. As with the Rand index, a higher adjusted Rand index means a higher correspondence between the two partitions. Please refer to our supplementary web site or (Yeung and Ruzzo, 2000) for a detailed description of the adjusted Rand index.

2.2 Subsets of PC's

Motivated by Chang's theoretical result (Chang, 1983), we would like to compare the effectiveness of clustering with the first few PC's to that of other sets of PC's. In particular, if there exists a set of "best" PC's that is most effective in capturing cluster structure, it would be interesting to compare the performance of this set of "best" PC's to the traditional wisdom of clustering with the first few PC's of the data. Since no such set of "best" PC's is known, we used the adjusted Rand index with the external criterion to determine if a set of PC's is effective in clustering. One way to determine the set of PC's that gives the maximum adjusted Rand index is by exhaustive search over all possible sets of PC's. However, exhaustive search is very computationally intensive. Therefore, we used heuristics to search for a set of PC's with high adjusted Rand index.

The greedy approach: A simple heuristic we implemented is the *greedy* approach, which is similar to the forward sequential search algorithm (Aha and Bankert, 1996). Let m_0 be the minimum number of PC's to be clustered, and p be the number of experimental conditions in the data.

- This approach starts with an exhaustive search for a set of m_0 PC's with maximum adjusted Rand index. Denote the optimum set of PC's as S_{m_0}.
- For each $m = (m_0 + 1), \ldots, p$,
  - For each component $currPC$ not in S_{m-1}
    - The data with all the genes projected onto components $S_{m-1} \cup \{currPC\}$ is clustered, and the adjusted Rand index is computed.
    - Record the maximum adjusted Rand index over all possible $currPC$.
  - S_m is the union of the component with maximum adjusted Rand index and S_{m-1}.

The modified greedy approach: The modified greedy approach requires an additional integer parameter, k, which represents the number of *best* solutions to keep in each search step. Denote the optimum k sets of components as $\mathcal{S}_m = \{S_m^1, \ldots, S_m^k\}$, where $m = m_0, \ldots, p$. This approach also starts with an exhaustive search for m_0 PC's with the maximum adjusted Rand index. However, k sets of components which achieve the top k adjusted Rand indices are stored. For each m (where $m = (m_0 + 1), \ldots, p$) and each of the S_m^i (where $i = 1, \ldots, k$), one additional component that is not already in S_{m-1}^i is added to the set of components, the subset of data with the extended set of components is clustered, and the adjusted Rand index is computed. The top k sets of m components that achieve the highest adjusted Rand indices are stored in $\mathcal{S}_m$. The modified greedy approach allows the search to have more choices in searching for a set of components that gives a high adjusted Rand index. Note that when $k = 1$, the modified greedy approach is identical to the simple greedy approach, and when $k = \binom{p}{m}$, the modified greedy approach is reduced to exhaustive search. So the choice for k is a tradeoff between running time and quality of solution. In our experiments, k is set to be 3.

2.3 Summary

Given a gene expression data set with n genes and p experimental conditions, our evaluation methodology consists of the following steps:

1. A clustering algorithm is applied to the given data set, and the adjusted Rand index with the external criterion is computed.
2. PCA is applied to the given data set. The same clustering algorithm is applied to the first m PC's (where $m = m_0, \ldots, p$). The adjusted Rand index is computed for each of the clustering results using the first m PC's.
3. The same clustering algorithm is applied to sets of PC's computed with the greedy and the modified greedy approaches.

2.4 Random PC's and Random Projections

As a control, we also investigated the effect on the quality of clusters obtained from random sets of PC's. Multiple sets of random PC's (30 in our experiments) were chosen to compute the average and standard deviation of the adjusted Rand indices.

We also compared the quality of clustering results from random PC's to that of random orthogonal projections of the data. Again, multiple sets (30) of random orthogonal projections were chosen to compute the average and standard deviations.

3 Data sets

We used two gene expression data sets with external criteria, and three sets of synthetic data to evaluate the effectiveness of PCA. The word *class* refers to a group in the external criterion that is used to assess clustering results. The word *cluster* refers to clusters obtained by a clustering algorithm. We assume both classes and clusters are partitions of the data, *i.e.*, every gene is assigned to exactly one class and to exactly one cluster.

3.1 Gene expression data sets

The ovary data: A subset of the ovary data obtained by (Schummer *et al.*, 1999) and (Schummer, 2000) is used. The ovary data set was generated by hybridizing to a membrane array containing a randomly selected cDNA library. The subset of the ovary data we used contains 235 clones and 24 tissue samples, 7 of which are derived from normal tissues, 4 from blood samples, and the remaining 13 from ovarian cancers in various stages of malignancy. The tissue samples are the experimental conditions. The 235 clones were sequenced, and discovered to correspond to 4 different genes. The numbers of clones corresponding to each of the four genes are 58, 88, 57, and 32 respectively. We expect clustering algorithms to separate the four different genes. Hence, the four genes form the four class external criterion for this data set. Different clones may have different hybridization intensities. Therefore, the data for each clone was normalized across the 24 experiments to have mean 0 and variance 1.

The yeast cell cycle data: The second gene expression data set we used is the yeast cell cycle data set (Cho *et al.*, 1998) which shows the fluctuation of expression levels of approximately 6000 genes over two cell cycles (17 time points). (Cho *et al.*, 1998) identified 420 genes which peak at different time points and categorized them into five phases of cell cycle. Out of the 420 genes they classified, 380 genes were classified into only one phase (some genes peak at more than one phase in the cell cycle). Since the 380 genes were identified according to the peak times of genes, we expect clustering results to correspond to the five phases to a certain degree. Hence, we used the 380 genes that belong to only one class (phase) as our external criterion. The data was normalized to have mean 0 and variance 1 across each cell cycle as suggested in (Tamayo *et al.*, 1999).

3.2 Synthetic data sets

Since the array technology is still in its infancy, the "real" data may be noisy, and clustering algorithms may not be able to extract all the classes contained in the data. There may also be information in real data that is not known to biologists. Therefore, we complemented our empirical 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. The following three sets of synthetic data represent our preliminary effort on synthetic gene expression data generation. We do not claim that any of the three synthetic data sets capture all of the characteristics of gene expression data. Each of the synthetic data set has strengths and weaknesses. By using *all* three sets of synthetic data, we hope to achieve a thorough comparison study capturing many different aspects of expression data.

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. We also ran experiments on larger synthetic data sets and observed similar results (see supplementary web site for details).

Mixture of normal distributions on the ovary data: Visual inspection of the ovary data suggests that the data is not too far from normal. Among other sources of variation, the expression levels for different clones of the same gene are not identical because the clones represent different portions of the cDNA. Figure 2(a) shows the distribution of the expression levels in a normal tissue in a class (gene) from the ovary data. We found that the distributions of the normal tissue samples are typically closer to normal distributions than those of tumor samples, for example, Figure 2(b).

The sample covariance matrix and the mean vector of each of the four classes (genes) in the ovary data are computed. Each class in the 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 original ovary data.

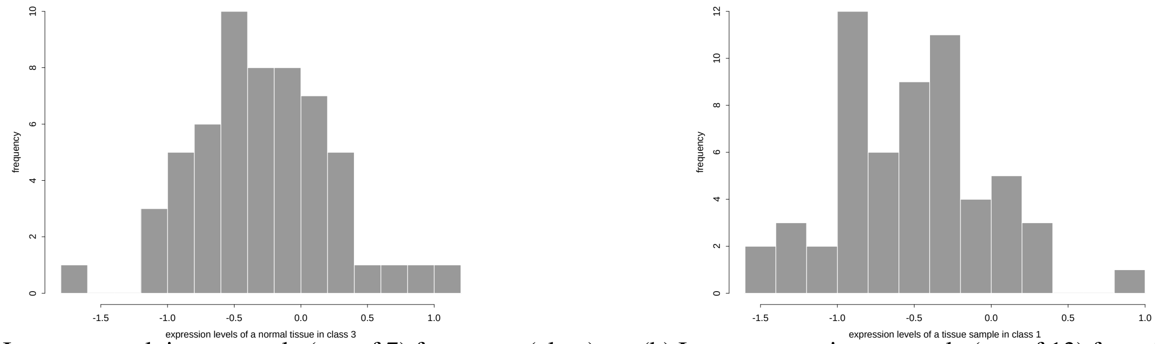
This synthetic data set preserves the covariance between the tissue samples in each gene. It also preserves the mean vectors of each class. The weakness of this synthetic data set is that the assumption of the underlying multivariate normal distribution for each class may not be true for real data.

Randomly resampled ovary data: In these data sets, the random data for an observation in class c (where $c = 1, \ldots, 4$) under experimental condition j (where $j = 1, \ldots, 24$) are 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 the ovary data.

This data set does not assume any underlying distribution. However, any possible correlation between tissue samples (for example, the normal tissue samples may be correlated) is not preserved due to the independent random sampling of the expression levels from each experimental condition. Hence, the resulting sample covariance matrix of this randomly resampled data set would be close to diagonal. However, inspection of the original ovary data shows that the sample covariance matrices are not too far from diagonal. Therefore, this set of randomly resampled data represents reasonable replicates of the original ovary data set.

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. There is evidence that the sine function correctly models the cell cycle behavior (see (Holter *et al.*, 2000) and (Alter *et al.*, 2000)). 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 $x_{i,j}$ be the simulated expression level of gene i and condition j in this data set with ten classes. Let $x_{i,j} = \delta_j + \lambda_j * (\alpha_i + \beta_i \phi(i,j))$, where $\phi(i,j) = \sin(\frac{2\pi j}{8} - w_{k(i)} + \epsilon)$ (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 8 time points (experiments).



(a) In one normal tissue sample (out of 7) for a gene (class) (b) In one tumor tissue sample (out of 13) for a different gene

Figure 2: Histograms of the distributions of the expression levels in the ovary data

$k(i)$ is the class number of gene i, which is chosen according to Zipf's Law (Zipf, 1949) to model classes with different sizes. Different classes are represented by different phase shifts $w_{k(i)}$, which are chosen according to the uniform distribution in the interval $[0, 2\pi]$. ϵ, which represents noise of gene synchronization, is chosen according to the standard normal distribution. λ_j is the amplitude control of condition j, and is chosen according to the normal distribution with mean 3 and standard deviation 0.5. δ_j, which represents an additive experimental error, is chosen according to the standard normal distribution. Each observation (row) is normalized to have mean 0 and variance 1 before PCA or any clustering algorithm is applied. A drawback of this model is the ad-hoc choice of the parameters for the distributions of α_i, β_i, λ_j, and δ_j.

4 Clustering algorithms and similarity metrics

We used three clustering algorithms in our empirical study: the *Cluster Affinity Search Technique* (CAST) (Ben-Dor and Yakhini, 1999), the hierarchical *average-link* algorithm, and the *k-means* algorithm (with average-link initialization) (Jain and Dubes, 1988). Please refer to our supplementary web site for details of the clustering algorithms. In our experiments, we evaluated the effectiveness of PCA on clustering analysis with both Euclidean distance and correlation coefficient, namely, CAST with correlation coefficient, average-link with both correlation and distance, and k-means with both correlation and distance. CAST with Euclidean distance usually does not converge, so it is not considered in our experiments. If Euclidean distance is used as the similarity metric, the minimum number of components in sets of PC's (m_0) considered is 2. If correlation is used, the minimum number of components (m_0) considered is 3 because there are at most 2 clusters if 2 components are used (when there are 2 components, the correlation coefficient is either 1 or -1).

5 Results and Discussion

Here are the overall conclusions from our empirical study:

- The quality of clustering results (*i.e.*, the adjusted Rand index with the external criterion) on the data after PCA is not necessarily higher than that on the original data on both real and synthetic data.
- We also showed that in most cases, the first m PC's (where $m = m_0, \ldots, p$) do not give the highest adjusted Rand index, *i.e.,* there exists another set of m components that achieves a higher adjusted Rand index than the first m components.
- There are no clear trends regarding the choice of the optimal number of PC's over all the data sets and over all the clustering algorithms and over the different similarity metrics. There is no obvious relationship between cluster quality and the number or set of PC's used.
- On average, the quality of clusters obtained by clustering random sets of PC's tend to be slightly lower than those obtained by clustering random sets of orthogonal projections, especially when the number of components is small.

In the following sections, the detailed experimental results are presented. In a typical result graph, the adjusted Rand index is plotted against the number of components. Usually the adjusted Rand index without PCA, the adjusted Rand index of the first m components, and the adjusted Rand indices using the greedy and modified greedy approaches are shown in each graph. Note that there is only one value for the adjusted Rand index computed with the original variables (without PCA), while the adjusted Rand indices computed using PC's vary with the number of components. Enlarged and colored versions of the graphs can be found on our supplementary web site. The results using the hierarchical average-link clustering algorithm turn out to show similar patterns to those using k-means (but with slightly lower adjusted Rand indices), and hence are not shown in this paper. The results of average-link can be found on our supplementary web site.

5.1 Gene expression data

5.1.1 The ovary data

CAST: Figure 3(a) shows the result on the ovary data using CAST as the clustering algorithm and correlation coefficient as the similarity metric. The adjusted Rand indices using the first m components (where $m = 3, \ldots, 24$) are mostly lower than those without PCA. However, the adjusted Rand indices using the greedy and modified greedy approaches for 4 to 22 components are higher than those without PCA. This shows that clustering with the first m PC's instead of the original variables may not help to extract the clusters in the data set, and that there exist sets of PC's (other than the first few which contain most of the variation in the data) that achieve higher adjusted Rand indices than clustering with the original variables. Moreover, the adjusted Rand indices computed using the greedy and modified greedy approaches are not very different. Figure 3(b) shows the additional results of the average adjusted Rand indices of random sets of PC's and random orthogonal projections. The standard deviation in the adjusted Rand indices of the multiple runs (30) of random orthogonal projections are represented by the error bars in Figure 3(b). The adjusted Rand indices of clusters from random sets of PC's are more than one standard deviation lower than those from random orthogonal projections when the number of components is small. Random sets of PC's have larger variations over multiple random runs, and their error bars overlap with those of the random orthogonal projections, and so are not shown for clarity of the figure. It turns out that Figure 3(b) shows typical behavior of random sets of PC's and random orthogonal projections over different clustering algorithms and similarity metrics, and hence those curves will not be shown in subsequent figures.

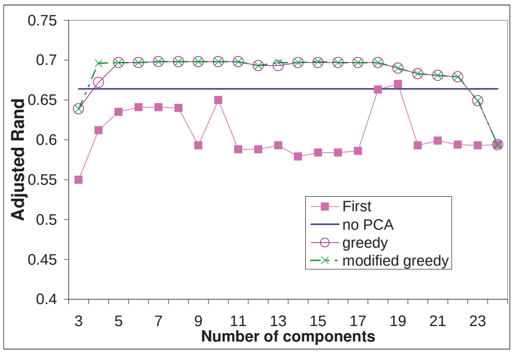
K-means: Figures 3(c) and 3(d) show the adjusted Rand indices using the k-means algorithm on the ovary data with correlation and Euclidean distance as similarity metrics respectively. Figure 3(c) shows that the adjusted Rand indices using the first m components tends to increase from below the index without PCA to above that without PCA as the number of components increases. However, the results using the same algorithm but Euclidean distance as the similarity metric show a very different picture (Figure 3(d)): the adjusted Rand indices are high for first 2 and 3 PC's and then drop drastically to below that without PCA. Manual inspection of the clustering result of the first 4 PC's using k-means with Euclidean distance shows that two classes are combined in the same cluster while the clustering result of the first 3 PC's separates the 4 classes, showing that the drastic drop in the adjusted Rand index reflects degradation of cluster quality with additional PC's. When the data points are visualized in the space of the first three PC's, the four classes are reasonably well-separated in the Euclidean space. However, when the data points are visualized in the space of the first, second and fourth PC's, the classes overlap. The addition of the fourth PC caused the cluster quality to drop. With both the greedy and the modified greedy approaches, the fourth PC was the second to last PC to be added. Therefore, we believe that the addition of the fourth PC makes the separation between classes less clear. Figures 3(c) and 3(d) show that different similarity metrics may have very different effect on clustering with PC's.

The adjusted Rand indices using the modified approach in Figure 3(c) show an irregular pattern. In some instances, the adjusted Rand index computed using the modified greedy approach is even lower than that using the first few components and that using the greedy approach. This shows, not surprisingly, that our heuristic assumption for the greedy approach is not always valid. Nevertheless, the greedy and modified greedy approaches show that there exists other sets of PC's that achieve higher adjusted Rand indices than the first few PC's most of the time.

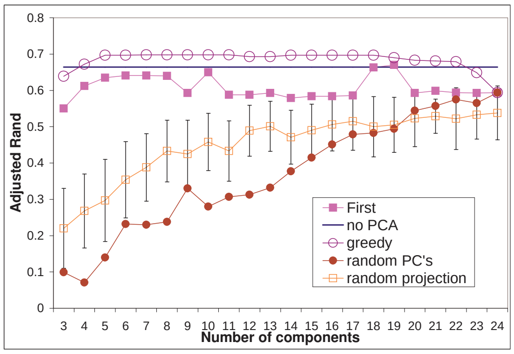
Effect of clustering algorithm: Note that the adjusted Rand index without PCA using CAST with correlation (0.664) is much higher than that using k-means (0.563) with the same similarity metric. Manual inspection of the clustering results without PCA shows that only CAST clusters mostly contain clones from each class, while k-means clustering results combine two classes into one cluster. This again confirms that higher adjusted Rand indices reflect higher cluster quality with respect to the external criteria. With the first m components, CAST with correlation has a similar range of adjusted Rand indices to the other algorithms (approximately between 0.55 to 0.68).

Choosing the number of first PC's: A common rule of thumb to choose the number of first PC's is to choose the smallest number of PC's such that a chosen percentage of total variation is exceeded. For the ovary data, the first 14 PC's cover 90% of the total variation in the data. If the first 14 PC's are chosen, it would have a detrimental effect on cluster quality if CAST with correlation, k-means with distance, or average-link with distance is the algorithm being used.

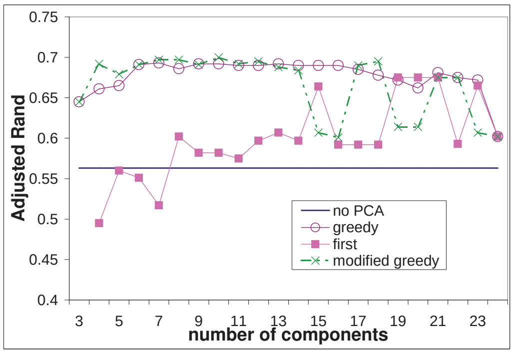
When correlation is used (Figures 3(a) and 3(c)), the adjusted Rand index using all 24 PC's is not the same as that using the original variables. On the other hand, when Euclidean distance is used (Figure 3(d)), the adjusted Rand index using all 24 PC's is the same as that with the original variables. This is because the Euclidean distance between a pair of genes using all the PC's is the same as that using the original variables. Correlation coefficients, however, are not preserved after PCA.



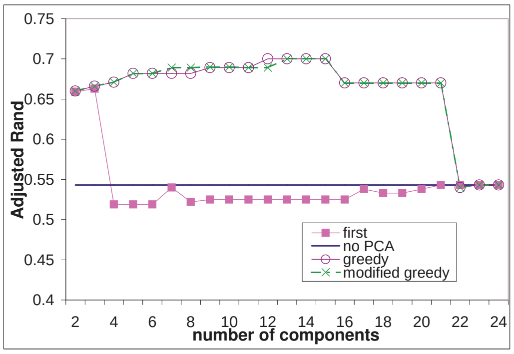
(a) CAST with correlation



(b) CAST with correlation

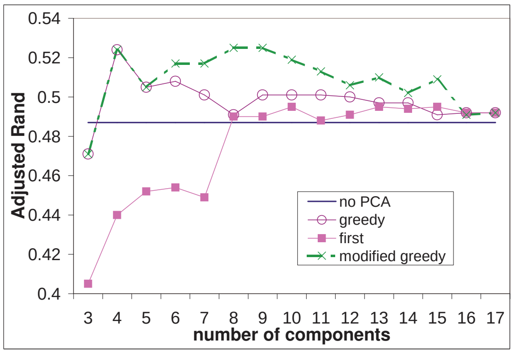


(c) k-means with correlation

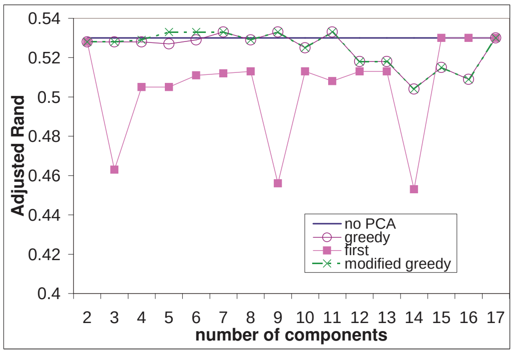


(d) k-means with Euclidean distance

Figure 3: Adjusted Rand index against the number of components on the ovary data.



(a) CAST with correlation



(b) k-means with Euclidean distance

Figure 4: Adjusted Rand index against the number of components on the yeast cell cycle data.

5.1.2 The yeast cell cycle data

CAST: Figure 4(a) shows the result on the yeast cell cycle data using CAST as the clustering algorithm and correlation coefficient as the similarity metric. The adjusted Rand indices using the first 3 to 7 components are lower than that without PCA, while the adjusted Rand indices with the first 8 to 17 components are comparable to that without PCA.

K-means: Figure 4(b) shows the result on the yeast cell cycle data using k-means with Euclidean distance. The adjusted Rand indices without PCA are relatively high compared to those using PC's. Figure 4(b) on the yeast cell cycle data shows a very different picture than Figure 3(d) on the ovary data. This shows that the effectiveness of clustering with PC's depends on the data set being used.

5.2 Synthetic data

5.2.1 Mixture of normal distributions on the ovary data

CAST: The results using this synthetic data set are similar to those of the ovary data in Section 5.1.1. Figure 5(a) shows the results of our experiments on the synthetic mixture of normal distributions on the ovary data using CAST as the clustering algorithm and correlation coefficient as the similarity metric. The lines in Figure 5(a) represent the average adjusted Rand indices over the 10 replicates of the synthetic data, and the error bars represent one standard deviation from the mean for the modified greedy approach and for using the first m PC's. The error bars show that the standard deviations using the modified greedy approach tend to be lower than that using the first m components. A careful study also shows that the modified greedy approach has lower standard deviations than the greedy approach (data not shown here). The error bars for the case without PCA are not shown for clarity of the figure. The standard deviation for the case without PCA is 0.064 for this set of synthetic data, which would overlap with those using the first components and the modified greedy approach. Using the Wilcoxon signed rank test (Hogg and Craig, 1978), we show that the adjusted Rand index without PCA is greater than that with the first m components at the 5% significance level for all $m = 3, \ldots, 21$. A manual study of the experimental results from each of the 10 replicates (details not shown here) shows that 8 out of the 10 replicates show very similar patterns to the average pattern in Figure 5(a), *i.e.,* most of the cluster results with the first m components have lower adjusted Rand indices than that without PCA, and the results using the greedy and modified greedy approach are slightly higher than that without PCA. In the following results, only the average patterns will be shown. Figure 5(a) shows a similar trend to real data in Figure 3(a), but the synthetic data has higher adjusted Rand indices for the clustering results without PCA and with the greedy and modified greedy approaches.

K-means: The average adjusted Rand indices using the k-means algorithm with the correlation and Euclidean distance as similarity metrics are shown in Figure 5(b) and Figure 5(c) respectively. In Figure 5(b), the average adjusted Rand indices using the first m components gradually increase as the number of components increases. Using the Wilcoxon signed rank test, we show that the adjusted Rand index without PCA is less than that with the first m components (where $m = 5, \ldots, 24$) at the 5% significance level. In Figure 5(c), the average adjusted Rand indices using the first m components are mostly below that without PCA. The results using average-link (not shown here) are similar to the results using k-means.

5.2.2 Randomly resampled ovary data

Figures 6(a) and 6(b) show the average adjusted Rand indices using CAST with correlation, and k-means with Euclidean distance on the randomly resampled ovary data. The general trend is very similar to the results on the ovary data and the mixture of normal distributions.

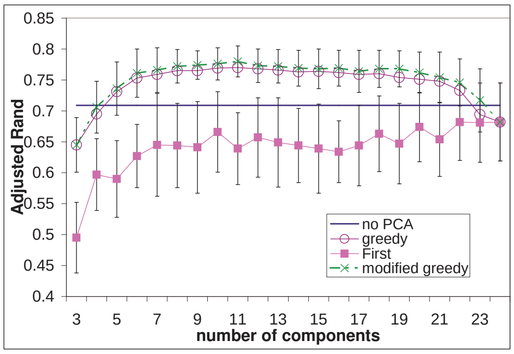
5.2.3 Cyclic data

Figure 7(a) shows the average adjusted Rand indices using CAST with correlation. The quality of clusters using the first PC's are worse than that without PCA, and is not very sensitive to the number of first PC's used.

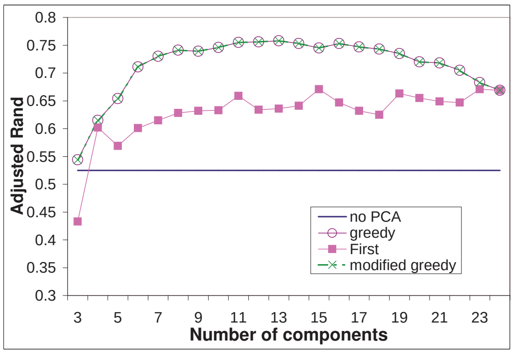
Figure 7(b) shows the average adjusted Rand indices with the k-means algorithm with Euclidean distance as the similarity metric. Again, the quality of clusters from clustering with the first PC's is not higher than that from clustering with the original variables.

5.3 Summary of results

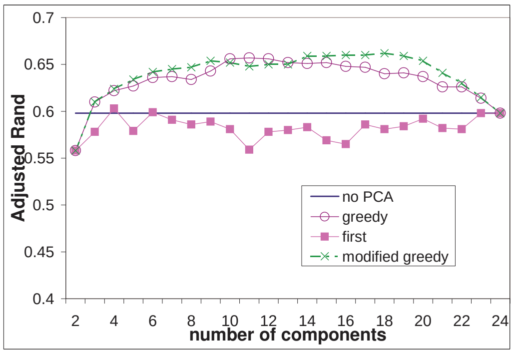
On both real and synthetic data sets, the adjusted Rand indices of clusters obtained using PC's determined by the greedy or modified greedy approach tend to be higher than the adjusted Rand index from clustering with the original variables. Table 1 summarizes the comparisons of the average adjusted Rand indices from clustering with the first PC's (averaged over the range of number of components) to the adjusted Rand indices from clustering the original real expression data. An entry is marked "+" in Table 1 if the average adjusted Rand index from clustering with the first components is higher than the adjusted Rand index from clustering the original data. Otherwise, an entry is marked with a "-". Table 1 shows that with the exception of k-means with correlation and average-link with correlation on the ovary data set, the average adjusted Rand indices using different numbers of the first components are lower than the adjusted Rand indices from clustering the



(a) CAST with correlation

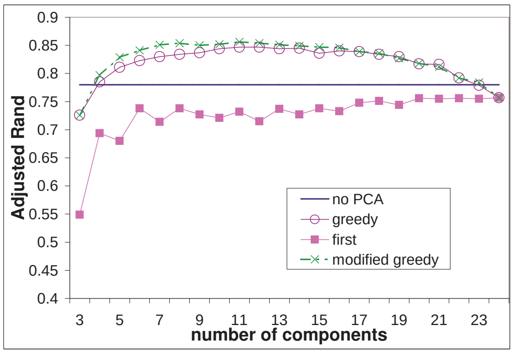


(b) k-means with correlation

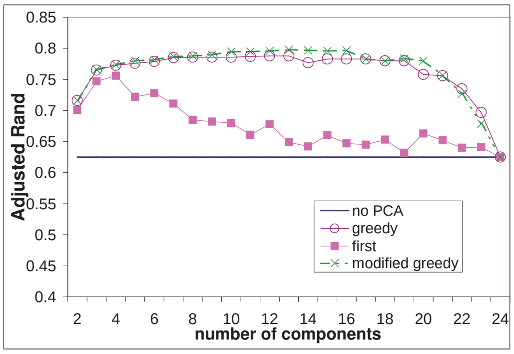


(c) k-means with Euclidean distance

Figure 5: Average adjusted Rand index against the number of components on the mixture of normal synthetic data.

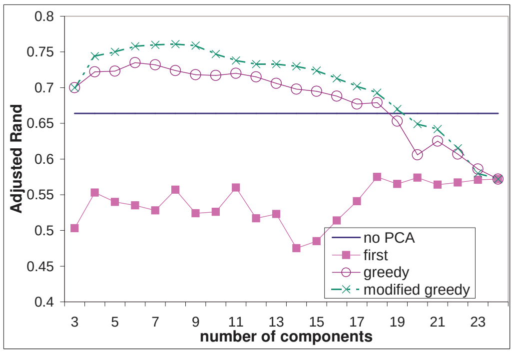


(a) CAST with correlation

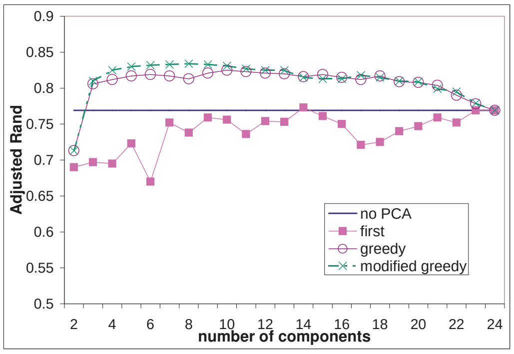


(b) k-means with Euclidean distance

Figure 6: Average adjusted Rand index against the number of components on the randomly resampled ovary data.



(a) CAST with correlation



(b) k-means with Euclidean distance

Figure 7: Average adjusted Rand index against the number of components on the cyclic data.

original data. On the synthetic data sets, we applied the one-sided Wilcoxon signed rank test to compare the adjusted Rand indices from clustering the first components to the adjusted Rand index from clustering the original data set on the 10 replicates. The p-values, averaged over the full range of possible numbers of the first components, are shown in Table 2. A low p-value suggests rejecting the null hypothesis that the adjusted Rand indices from clustering with and without PCA are comparable. Table 2 shows that the adjusted Rand indices from the first components are significantly lower than those from without PCA on both the mixture of normal and the cyclic synthetic data sets when CAST with correlation is used. On the other hand, the adjusted Rand indices from the first components are significantly higher than those from without PCA when k-means with correlation is used on the mixture of normal synthetic data or when average-link with correlation is used on the randomly resampled data. However, the latter results are not clear successes for PCA since (1) they assume that the correct number of classes is known (which would not be true in practice), and (2) CAST with correlation gives better results on the original data sets without PCA in both cases. The average p-values of k-means with correlation on the cyclic data are not available because the iterative k-means algorithm does not converge on the cyclic data sets when correlation is used as the similarity metric.

6 Conclusions

Our experiments on two real gene expression data sets and three sets of synthetic data show that clustering with the PC's instead of the original variables does not necessarily improve, and may worsen, cluster quality. Our empirical study shows that the traditional wisdom that the first few PC's (which contain most of the variation in the data) may help to extract cluster structure is generally *not* true. We also show that there usually exists some other sets of m PC's that achieve higher quality of clustering results than the first m PC's.

Our empirical results show that clustering with PC's has different impact on different algorithms and different similarity metrics (see Table 1 and Table 2). When CAST is used with correlation as the similarity metric, clustering with the first m PC's gives a lower adjusted Rand index than clustering with the original variables for most of $m = 3, \ldots, 24$, and this is true in both real and synthetic data sets. On the other hand, when k-means is used with correlation as the similarity metric, using *all* of the PC's in cluster analysis instead of the original variables usually gives higher or similar adjusted Rand indices on all of our real and synthetic data sets. When Euclidean distance is used as the similarity metric on the ovary data or the synthetic data sets based on the ovary data, clustering (either with k-means or average-link) using the first few PC's usually achieves higher or comparable adjusted Rand indices to without PCA, but the adjusted Rand indices drop sharply with more PC's. Since the Euclidean distance computed with the first m PC's is just an approximation to the Euclidean distance computed with all the experiments, the first few PC's probably contain most of the cluster information while the last PC's are mostly noise. There is no clear indication from our results of how many PC's to use in the case of Euclidean distance. Choosing PC's by the rule of thumb to cover 90% of the total variation in the data are too many in the case of Euclidean distance on the ovary data and yeast cell cycle data. Based on our empirical results, we recommend against using the first few PC's if CAST with correlation is used to cluster a gene expression data set. On the other hand, we recommend using all of the PC's if k-means with correlation is used instead. However, the increased adjusted Rand indices using the "appropriate" PC's with k-means and average-link are comparable to that of CAST using the original variables in many of our results. Therefore, choosing a good clustering algorithm is as important as choosing the "appropriate" PC's.

There does not seem to be any general relationship between cluster quality and the number of PC's used based on the results on both real and synthetic data sets. The choice of the first few components is usually not optimal (except when Euclidean distance is used), and often achieves lower adjusted Rand indices than without PCA. There usually exists another set of PC's (determined by the greedy or modified greedy approach) that achieves higher adjusted Rand indices than clustering with the original variables or with the first m PC's. However, both the greedy and the modified greedy approaches require the external criteria to determine a "good" set of PC's. In practice, external criteria are seldom available for gene expression data, and so we cannot use the greedy or the modified greedy approach to choose a set of PC's that captures the cluster structure. Moreover, there does not seem to be any general trend for the the set of PC's chosen by the greedy or modified greedy approach that achieves a high adjusted Rand index. A careful manual inspection of our empirical results shows that the first two PC's are usually chosen in the exhaustive search step for the set of m_0 components that give the highest adjusted Rand indices. In fact, when CAST is used with correlation as the similarity metric, the 3 components found in the exhaustive search step *always* include the first two PC's on *all* of our real and synthetic data sets. The first two PC's are *usually* returned by the exhaustive search step when k-means with correlation, or k-means with Euclidean distance, or average-link with correlation is used. We also tried to generate a set of random PC's that always includes the first two PC's, and then apply clustering algorithms and compute the adjusted Rand indices. The result is that the adjusted Rand indices are similar to that computed using the first components.

To conclude, our empirical study shows that clustering with the PC's enhances cluster quality only when the right number

data	*CAST correlation*	*k-means correlation*	*k-means distance*	*average-link correlation*	*average-link distance*
ovary data	−	+	−	+	−
cell cycle data	−	−	−	−	−

Table 1: Comparisons of the average adjusted Rand indices from clustering with different numbers of the first components to the adjusted Rand indices from clustering the original real expression data. An entry marked "+" indicates the average quality from clustering with the first components is higher than that from clustering the original data.

synthetic data	alternative hypothesis	*CAST correlation*	*k-means correlation*	*k-means distance*	*average-link correlation*	*average-link distance*
mixture of normal	no PCA $>$ first	**0.039**	0.995	0.268	0.929	0.609
mixture of normal	no PCA $<$ first	0.969	**0.031**	0.760	0.080	0.418
randomly resampled	no PCA $>$ first	0.243	0.909	0.824	0.955	0.684
randomly resampled	no PCA $<$ first	0.781	0.103	0.200	**0.049**	0.337
cyclic data	no PCA $>$ first	**0.023**	not available	0.296	0.053	0.799
cyclic data	no PCA $<$ first	0.983		0.732	0.956	0.220

Table 2: Average p-value of the Wilcoxon signed rank test over different number of components on synthetic data sets. Average p-values below 5% are bold faced.

of components or when the right set of PC's is chosen. However, there is not yet a satisfactory methodology to determine the number of components or an informative set of PC's without relying on external criteria of the data sets. Therefore, in general, we recommend against using PCA to reduce dimensionality of the data before applying clustering algorithms unless external information is available. Moreover, even though PCA is a great tool to reduce dimensionality of gene expression data sets for visualization, we recommend cautious interpretation of any cluster structure observed in the reduced dimensional subspace of the PC's. We believe that our empirical study is one step forward to investigate the effectiveness of clustering with the PC's instead of the original variables.

Acknowledgement

We would like to thank Michèl Schummer from the Institute of Systems Biology for the ovary data set and his feedback. We would also like to thank Mark Campbell, Pedro Domingos, Chris Fraley, Phil Green, David Haynor, Adrian Raftery, Rimli Sengupta, Jeremy Tantrum and Martin Tompa for their feedback and suggestions. This work is partially supported by NSF grant DBI-9974498.

References

Aha, D. W. and Bankert, R. L. (1996) A comparative evaluation of sequential feature selection algorithms. In Fisher, D. and Lenz, J. H. (eds.), *Artificial Intelligence and Statistics V*, New York: Springer-Verlag, 199–206.

Alter, O., Brown, P. O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Science USA*, **97**, 10101–10106.

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, 33–42.

Chang, W. C. (1983) On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, **32**, 267–275.

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. *Molecular Cell*, **2**, 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

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

Hogg, R. V. and Craig, A. T. (1978) *Introduction to Mathematical Statistics*. 4th edn., Macmillan Publishing Co., Inc., New York, NY.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. and Fedoroff, N. V. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Science USA*, **97**, 8409–8414.

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

Jain, A. K. and Dubes, R. C. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

Jolliffe, I. T. (1986) *Principal Component Analysis*. New York : Springer-Verlag.

Jolliffe, I. T., Jones, B. and Morgan, B. J. T. (1980) cluster analysis of the elderly at home: a case study. *Data analysis and Informatics*, 745–757.

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

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

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

Raychaudhuri, S., Stuart, J. M. and Altman, R. B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, Hawaii, 452–463.

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) Comparative hybridization of an array of 21500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Genes*, **238**, 375–385.

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. *Proceedings of the National Academy of Science USA*, **96**, 2907–2912.

Yeung, K. Y. and Ruzzo, W. L. (2000) An empirical study on principal component analysis for clustering gene expression data. Tech. Rep. UW-CSE-00-11-03, Dept. of Computer Science and Engineering, University of Washington.

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

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