Cluster Analysis of Gene Expression Data

Ka Yee Yeung

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2001

Program Authorized to Offer Degree: Department of Computer Science and Engineering

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Ka Yee Yeung

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

_____

Walter L. Ruzzo

Reading Committee:

_____

David R. Haynor

_____

Walter L. Ruzzo

_____

Martin Tompa

Date: _____

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature_____

Date_____

University of Washington

Abstract

Cluster Analysis of Gene Expression Data

by Ka Yee Yeung

Chair of Supervisory Committee:

Professor Walter L. Ruzzo
Department of Computer Science and Engineering

The invention of DNA microarrays allows us to study simultaneous variations of genes at the genome-wide scale. A typical gene expression data set consists of thousands or even tens of thousands of genes, and a few dozens experiments. Cluster analysis is the art of finding groups in a given data set such that objects in the same group are similar to each other while objects in different groups are dissimilar. There are many applications for clustering gene expression data.

Many different clustering algorithms and analytical techniques have been applied to gene expression data. Success of various analytical methodologies in specific instances has been reported, but extensive quantitative evaluations of clustering methodologies are rare. Since different analytical approaches may produce different clustering results, there is a great need to evaluate clustering techniques in order to choose an appropriate approach. An underlying theme of this dissertation is systematic evaluations of clustering methodologies on gene expression data. Specifically, we proposed a data-driven methodology, called the figure of merit (FOM) methodology, to compare the quality of clusters from heuristic-based clustering algorithms. We also showed that the model-based clustering approach, which assumes the Gaussian mixture model, produces relatively high quality clusters. The probabilistic framework in the model-based approach allows us to infer the correct number of clusters, and to compare different models. Moreover, we investigated the effectiveness

of a dimension reduction technique called principal component analysis as a pre-processing step before cluster analysis.

Our main contributions are evaluation methodologies of analytical techniques in clustering gene expression data. We employed an external validation approach, which evaluates clustering results by comparing to external prior knowledge of the data, to assess the performance of internal validation approaches, which do not require any external knowledge of the data. In particular, we showed that our FOM methodology and the model-based approach, which do not require any external knowledge of the data, produce comparisons of clustering algorithms that are consistent with comparisons to external knowledge. Since external knowledge is seldom available for gene expression data, our work provides practical evaluation frameworks for assessing clustering results on gene expression data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ADENOCARCINOMA: a malignant tumor.[1]

CDNA: DNA synthesized from an RNA template using reverse transcriptase.[2]

CYTOKERATIN: intermediate filament proteins of epithelial cells.

DNA: deoxyribonucleic acid; the molecule that encodes genetic information.

ENDOSCOPIC BIOPSY: tissue sample taken from the esophagus or stomach during endoscopy which can be used for genetic testing or histologic evaluation.[1]

ENDOSCOPY: a procedure in which a flexible tube with a fiber optic camera is inserted into the esophagus and stomach in order to visualize any abnormalities and collect tissue samples for further analysis.[1]

EPITHELIUM: membrane tissue composed of one or more layers of cells separated by very little intercellular substance and forming the covering of most internal and external surfaces of the body and its organs.[1]

ESOPHAGECTOMY: surgical removal of the esophagus.[3]

GASTROESOPHAGEAL REFLUX DISEASE (GERD): usually referred to as "heartburn"; a burning discomfort in the chest and regurgitation of sour tasting gastric juice into the mouth are classic symptoms of GERD.[3]

---

[1]From the web page of the Seattle Barrett's Esophagus Program at http://www.fhcrc.org/science/phs/barretts/glossary.htm

[2]From the BioTech Life Sciences Resources and Reference Tools at http://biotech.icmb.utexas.edu

[3]From http://www.barrettsinfo.com

GENE: a segment of DNA which normally specifies a functional unit.

GENE EXPRESSION: the process by which a gene's coded information is converted to the structures present and operating in the cell.[2]

HISTOLOGY: the anatomical study of the microscopic structure of animal and plant tissues.[1]

IN VIVO: a biological or biochemical process occurring within a living organism.[2]

METAPLASIA: abnormal replacement of cells of one type by cells of another.[4]

NEOPLASTIC: an abnormal new growth of tissue in animals or plants.[1]

PHENOTYPE: the physical appearance or observable characteristics of an organism.[2]

PREMALIGNANT: preceding the development of cancer.[1]

PROBE: a DNA or RNA fragment which has been labeled; usually used to identify DNA or RNA sequences which are closely related in sequence.

RNA: ribonucleic acid; a chemical found in the nucleus and cytoplasm of cells. The structure of RNA is similar to that of DNA.[2]

REVERSE TRANSCRIPTASE: an enzyme that synthesizes a cDNA strand from an RNA template.

TRANSCRIPTION: the synthesis of RNA from DNA.

---

[4]Merriam Webster Medical Dictionary

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to her PhD advisor, Professor Larry Ruzzo, for his patience and guidance. She would also like to acknowledge many other colleagues and friends who helped to shape her research: Dr. Mike Barrett, Jeremy Buhler, Dr. Jeff Delrow, Chris Fraley, Professor David Haynor, Trey Ideker, Professor Dick Karp, Professor Adrian Raftery, Dr. Brian Reid, Dr. Michèl Schummer, and Professor Martin Tompa. Finally, this thesis work would not be possible without the love and encouragement from her parents, her sister, and her fiancé.

Chapter 1

INTRODUCTION

Even until about a decade ago, being able to study simultaneous variations of genes at the genome-wide scale was only a dream, and happened only in science fiction. With the invention of the DNA microarrays, real life technology has caught up with science fiction, and measuring variations of genes at the genome-wide scale is no longer a dream. DNA microarrays offer a global view on the levels of activity of many genes simultaneously. In fact, Lander [49] suggested that the global views provided by DNA microarrays would help us to understand the "Periodic Table of Life". Massive amounts of DNA microarray data have been generated by many researchers. The challenge is the development of analytical techniques to make sense out of the large amounts of biological data.

With current technology, the expression levels of thousands of genes can be measured from a single DNA microarray (sometimes called a DNA chip). The number of measurements of gene expression levels on a single array is increasing with advances in technology. There are different technologies producing gene expression data. Two examples are the cDNA arrays in which the expression levels are measured with respect to a reference target [54] and the commercially available Affymetrix chips [13]. Shamir and Sharan [75] provided a concise summary of the different types of DNA microarrays. The types of array technology do not affect the analysis described in this dissertation unless otherwise stated. We can study the variations of thousands of genes under different experimental conditions using a series of DNA microarrays. The experimental condition for each DNA microarray is determined by the goal of the scientific study. For example, the experimental conditions represent different time points in Cho *et al.* [20], in which the variations of the yeast genes over the time course of two cell cycles are studied. In Schummer *et al.* [72], the goal is to

characterize genes with different expression levels in normal ovary tissue samples and cancerous ovary tissue samples, and so the experimental conditions represent different tissue samples.

In a typical gene expression data set, the number of genes is usually much larger than the number of experiments. Even a simple organism like yeast has approximately six thousand genes. It is estimated that humans have approximately thirty thousand to forty thousand genes [22]. Since the cost of the study is driven by the number of DNA chips used (and hence by the number of experiments), the number of experiments is mostly a few dozens in published gene expression papers. The number of experiments is expected to increase as the costs of DNA chips go down.

1.1 Clustering gene expression data

Cluster analysis is the art of finding groups in a given data set such that objects in the same group are similar to each other while objects in different groups are as dissimilar as possible [46]. Clustering is a very well-studied problem, and there are many algorithms designed for cluster analysis in the literature. 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.

There are many applications of clustering gene expression data. Genes with related functions are expected to have similar expression patterns, so clustering of genes may suggest possible roles for genes with unknown functions based on the known functions of some other genes that are placed in the same cluster. For example, Chu *et al.* [21] applied clustering to yeast genes to identify genes whose expression levels peak at different phases in sporulation. Clustering of genes is sometimes used as a preprocessing step in inferring regulatory networks. For example, Chen *et al.* [18] used clustering to identify genes that have similar expression patterns to reduce the size of the regulatory network to be inferred. Gene clusters from expression data can also be used with sequence data to identify upstream DNA sequence patterns specific to each expression cluster [78]. These upstream DNA sequence patterns may co-regulate genes within the same clusters. There are many published gene

expression data sets with different types of tissue samples. For example, some of the experiments may represent normal tissue samples while some experiments represent cancerous tissue samples at different stages of malignancy. Clustering the experiments may shed light on new subtypes of cancer which subsequently may require different treatment. For example, Golub *et al.* [32] demonstrated that clustering of the experiments can potentially lead to discovery of subtypes of leukemia.

Many clustering algorithms have been proposed for gene expression data. For example, Eisen *et al.* [27] applied a variant of the hierarchical average-link clustering algorithm to identify groups of co-regulated yeast genes. Ben-Dor *et al.* [10],[9] reported success with their CAST algorithm. The classic iterative k-means algorithm is also widely used to cluster gene expression data, for example, [78]. Tamayo *et al.* [77] used self-organizing maps (SOM) to identify gene clusters in the yeast cell cycle and human hematopoietic differentiation data sets. These algorithms can be applied to cluster either the genes or the experiments. Some of these algorithms will be discussed in more detail in Chapter 2. There are also clustering approaches that cluster both the genes and the experiments simultaneously, for example, Lazzeroni and Owen [50], Cheng and Church [19]. One of the applications of clustering the genes and the experiments simultaneously is to find subsets of genes with similar expression patterns with respect to subsets of experiments. In this dissertation, the focus is on clustering of genes, and the term *clustering* refers to clustering of the genes unless otherwise stated.

1.2 What this thesis is about

Many different analytical techniques have been used in the context of clustering gene expression data, and instances of success from many different methods have been reported in specific applications. Different analytical techniques usually lead to different results. There is little or no systematic comparison of different analytical approaches in the context of clustering gene expression data. Our goal is to propose methodologies for systematic evaluation of analytical techniques in clustering gene expression data. An underlying theme of this dissertation is application of new or existing methodologies to clustering gene expression data,

and quantitative evaluation of the methodologies. Specifically, we used gene expression data sets with external criteria to evaluate analytical methodologies by comparing clustering results to the external criterion (which serves as ideal clusters of the data). Chapter 2 covers the background for the rest of the dissertation. In particular, the clustering algorithms we applied, and both real gene expression data sets and synthetic data sets we used will be described. The statistic we used to assess agreement of clustering results to the external criterion is also described in Chapter 2.

Many different clustering algorithms have been proposed to analyze gene expression data, and success in their applications had been reported. However, there is no clustering algorithm of choice in the gene expression analysis community. Moreover, different clustering algorithms can potentially generate very different clusters on the same data set. A biologist with a gene expression data set is faced with the problem of choosing an appropriate clustering algorithm for his or her data set. In much of the published clustering work on gene expression, the success of clustering algorithms is assessed by visual inspection using biological knowledge (for example, Michaels *et al.* [61] and Eisen *et al.* [27]). Our FOM (*figure of merit*) methodology on comparing the performance of clustering algorithms (described in Chapter 3) provides a quantitative data-driven framework (which does not require any external prior knowledge of the data) to help biologists to choose a good clustering algorithm.

Other analytical techniques, such as principal component analysis (PCA), have also been proposed to analyze gene expression data. PCA [43] is a classical technique to reduce the dimensionality of the data set by transforming to a new set of variables to summarize the features of the data. 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.* [21] identified seven clusters in the sporulation data set, but Raychaudhuri *et al.* [67] suggested that there are no clusters present in the same data set by viewing the data points in the space of the first two principal components (PC's). There is a great need to investigate the effectiveness of PC's in capturing cluster structure on gene expression data. In particular, the effectiveness of the traditional approach of using the first few PC's (which capture most of the variation in the data) should be investigated. Our work described in

Chapter 4 is an attempt at such an empirical study.

Most of the clustering algorithms proposed to analyze gene expression data sets are based largely on heuristics. Clustering algorithms based on probability models offer a rigorous alternative to heuristic-based algorithms. In particular, model-based clustering assumes that the data is generated by a mixture of multivariate normal distributions. This Gaussian mixture model has been shown to be a powerful tool for many applications. The issues of selecting a "good" clustering method and determining the "correct" number of clusters are reduced to model selection problems under the probability framework. We evaluated both the quality of clusters and the model chosen by the model-based clustering approach using real and synthetic data sets with external criteria. Since we do not expect raw gene expression data to satisfy the Gaussian mixture assumption (which is implicit in the model-based approach), we also investigated the degree to which different transformations of real gene expression data sets satisfy the Gaussian mixture assumption. In addition, we also compared the results from the model-based approach to a leading heuristic clustering algorithm. The results are presented in Chapter 5.

The Barrett's esophagus data set [7] provides us with a test bed for some of our analytical techniques. Our collaboration with the Reid Lab at the Fred Hutchinson Cancer Research Center in Seattle provides us with biological feedback for our quantitative methodologies. Barrett's esophagus is a pre-cancer condition in which the normal squamous epithelium in the esophagus is replaced by the Barrett's epithelium. The Barrett's esophagus data set consists of different tissue types from the human gastrointestinal (GI) tract. In particular, there are three types of normal GI tissue samples (squamous, gastric and duodenum), and one pre-cancer tissue type (Barrett's epithelium). The goal of the study is to compare the pre-cancerous Barrett's epithelium to the surrounding normal GI tissues, and to study the expression levels of genes with respect to different tissue types. In order to reduce the number of genes in the clustering step, we developed a novel filtering approach to identify genes that are differentially expressed in each tissue type. We then applied our filtering approach and our FOM methodology to the filtered Barrett's esophagus data set. From our clustering results, our collaborators identified a set of biologically interesting genes. The data set and our analysis will be presented in Chapter 6.

1.3 Our Contributions

Our main contributions are evaluation methodologies of analytical techniques in clustering gene expression data. Data sets with external criteria enable us to investigate the performance of analytical techniques systematically. For example, we evaluated the clustering results using the PC's by comparing to the external criteria in Chapter 4. Since external criteria are rarely available for real gene expression data, we also used the external criteria to evaluate cluster validation approaches that do not require any external knowledge of the data. In Chapter 3 and Chapter 5, we demonstrated that the FOM methodology and the model-based approach produce comparisons of different clustering algorithms that are consistent with those from the external criteria. Our results provided confidence in the conclusions drawn from the FOM and model-based approaches in which no external knowledge of the data is required. In addition, we also provided specific guidelines and recommendations for clustering gene expression data in this dissertation. For example, we concluded that CAST and k-means tend to produce higher quality clusters than the hierarchical clustering approaches in Chapter 3, and we recommended clustering with the original data instead of using the PC's in Chapter 4.

Chapter 2

BACKGROUND

This chapter covers concepts that are used throughout this dissertation, including detailed description of clustering algorithms we used in our study, real and synthetic gene expression data sets to which we applied our analytical techniques, and the statistic used to assess agreement between clustering results and the external criteria (which serve as ideal clusters) of the data sets.

2.1 Cluster Analysis

2.1.1 Mathematical Formulations

Given a set of n objects $S = \{O_1, \ldots, O_n\}$, let $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ be a *partition* of S, *i.e.*, a set of subsets of S such that $\cup_{i=1}^{k} C_i = S$ and $C_i \cap C_j = \emptyset$ for $1 \leq i \neq j \leq k$. Each subset C_i (where $1 \leq i \leq k$) is called a *cluster*, and $\mathcal{C}$ is called a *clustering result*.[1] The goal of cluster analysis is to assign objects to clusters such that objects in the same cluster are more similar to each other while objects in different clusters are as dissimilar as possible. There are many ways to mathematically formulate the objectives of within-cluster homogeneity and between-cluster separation, leading to many different optimization problems.

A data set containing objects to be clustered is usually represented in one of two formats: the *data matrix* and the *similarity (or dissimilarity) matrix*. In a data matrix, the rows usually represent objects, and the columns usually represent features or attributes of the objects. Suppose there are n objects and p attributes. We assume the rows represent genes and the columns represent experiments, such that entry (i, e) in the data matrix D represents the expression level of gene i under experiment e, where $1 \leq i \leq n$ and $1 \leq e \leq p$. The ith row in the data matrix D (where $1 \leq i \leq n$), D_i, represents the *expression vector* of

[1] We assume *hard clustering* in this dissertation, *i.e.*, each object is assigned to one and only one cluster.

gene i across all p experiments. In clustering genes, the objects to be clustered are the genes. The similarity (or dissimilarity) matrix contains the pairwise similarity (or dissimilarity) of genes. Specifically, entry (i, j) in the similarity (or dissimilarity) matrix Sim represents the similarity (or dissimilarity) of gene i and gene j, where $1 \leq i, j \leq n$. The similarity (or dissimilarity) of gene i and gene j can be computed using the expression vectors of gene i and gene j from the data matrix. Hence, the similarity (or dissimilarity) matrix Sim can be computed from the data matrix D. However, the data matrix D cannot be fully recovered from the similarity matrix (especially when the number of experiments p is not known).

2.1.2 Similarity metrics

The measure used to compute similarity or dissimilarity between a pair of objects is called a *similarity metric*. Many different similarity metrics have been used in clustering gene expression data, among which the two most popular similarity metrics are correlation coefficient and Euclidean distance. Correlation coefficient is a similarity measure (a high correlation coefficient implies high similarity) while Euclidean distance is a dissimilarity measure (a high Euclidean distance implies low similarity).

The correlation coefficient between a pair of genes i and j $(1 \leq i, j \leq n)$ is defined as $\sum_{e=1}^{p}(D(i,e) - \mu_i)(D(j,e) - \mu_j)/(\parallel D_i \parallel \parallel D_j \parallel)$, where $\mu_i = \sum_{e=1}^{p} D(i,e)/p$ is the average expression level of gene i over all p experiments and $\parallel D_i \parallel = \sqrt{\sum_{e=1}^{p}(D(i,e) - \mu_i)^2}$ is the norm of the centered expression vector D_i. Correlation coefficients range from -1 to 1. The correlation coefficient of two genes with identical expression vectors is 1. Two genes having correlation coefficient 0 are said to be *uncorrelated*, and two genes having correlation coefficient -1 are said to be *anti-correlated*. Geometrically, correlation coefficients capture the *patterns* of expression levels of two genes. For example, two genes with different average expression levels but with expression levels peaking at the same experiments have a high correlation coefficient.

The Euclidean distance between a pair of genes i and j $(1 \leq i, j \leq n)$ is defined as $\sqrt{\sum_{e=1}^{p}(D(i,e) - D(j,e))^2}$. Euclidean distances are at least 0. A high Euclidean distance between a pair of genes indicates low similarity between the genes.

2.1.3 Clustering algorithms

There is a rich literature in clustering algorithms, and there are many different classifications of clustering algorithms. One classification is *model-based* versus *heuristic-based* clustering algorithms. The objects to be clustered are assumed to be generated from an underlying probability framework in the model-based clustering algorithms. The model-based approach will be discussed in details in Chapter 5. In the heuristic-based approach, an underlying probability framework is not assumed. In this chapter, heuristic-based clustering algorithms are discussed.

Another classification of clustering algorithms is *hierarchical* versus *partitional* clustering algorithms. In hierarchical clustering algorithms, objects are related by a tree structure called *dendrogram* such that objects in the same subtree are more similar to each other than objects in different subtrees. On the other hand, in *partitional* clustering algorithms, clusters are "flat", and there is no tree structure relating clusters. We implemented four hierarchical clustering algorithms (single-link, average-link, centroid-link and complete-link), and two partitional clustering algorithms (the *Cluster Affinity Search Technique* (CAST) [10], [9], and the *k-means* algorithm [55]) in our studies.

2.1.4 Hierarchical Algorithms

In agglomerative hierarchical algorithms, clusters are built bottom up. Hierarchical algorithms define a *dendrogram* (tree) relating objects (genes) in different subtrees. Initially, the dendrogram is empty (no objects are related by the dendrogram), and each object is in its own cluster, so the current number of clusters is the number of objects, n. In each step, the two clusters with the maximum *cluster similarity* (to be defined later) are merged to form a subtree, and hence the current number of clusters is reduced by 1. The two merged clusters are now connected by the same subtree in the dendrogram. This merging process is repeated until the desired number of clusters, k, is produced. Hence, there are k subtrees when the iterative merging process stops. Please refer to Hartigan [35] or Everitt [28] for a detailed description of the hierarchical clustering algorithms. The inputs to hierarchical clustering algorithms include the similarity matrix, *Sim* (which is used to compute cluster

similarity), and the desired number of clusters, k. In order to compare the clustering results from hierarchical clustering algorithms to those from partitional algorithms, we obtained k clusters by assigning each of the k subtrees to a cluster, and ignored the dendrogram in our studies.

Different cluster similarity criteria yield different clustering algorithms. The cluster similarity criteria of four popular hierarchical clustering algorithms (single-link, average-link, centroid-link and complete-link) will be discussed. Hierarchical clustering is very popular in clustering gene expression data. Eisen *et al.* [27] developed a software package for clustering gene expression data and visualizing clustering results. They used the hierarchical centroid-link algorithm in their implementations.[2]

In the hierarchical single-link clustering algorithm, the cluster similarity of two clusters is the maximum similarity between a pair of genes, one from each of the two clusters. Mathematically, the cluster similarity of clusters C_i and C_j ($1 \leq i \neq j \leq$ current number of clusters) for the hierarchical single-link algorithm is $\max_{x \in C_i, y \in C_j} Sim(x, y)$ assuming that $Sim(x, y)$ represents the pairwise similarity of object x and object y. Similarly, if Sim represents a dissimilarity measure, the cluster similarity is $\min_{x \in C_i, y \in C_j} Sim(x, y)$. It can be shown that the single-link criterion is closely related to the minimum spanning tree problem [33], which can be solved in time $O(n^2)$, where n is the number of genes. However, since two most similar objects are used to determine cluster similarity in each merging step, single-link can potentially cause *chaining* of clusters, in which clusters take geometrically elongated shapes, and two objects in the same cluster may have very low similarity.

With the complete-link criterion, the cluster similarity of two clusters is the minimum similarity between a pair of genes, one from each of the two clusters. Specifically, the cluster similarity of clusters C_i and C_j ($1 \leq i \neq j \leq$ current number of clusters) for the hierarchical complete-link algorithm is $\min_{x \in C_i, y \in C_j} Sim(x, y)$ assuming that Sim is a similarity measure. In other words, the two least similar objects from the two clusters are used to determine cluster similarity in each merging step. Assuming that the similarity matrix is given as input and stored in memory (hence using $O(n^2)$ space), the complete-link

[2] We would like to thank Sonia Leach for clarifying this issue.

algorithm has a running time of $O(n^2)$ [25].

With the average-link criterion, the cluster similarity of two clusters is the average pairwise similarity between genes in the two clusters. Specifically, the cluster similarity of clusters C_i and C_j ($1 \leq i \neq j \leq$ current number of clusters) for the hierarchical average-link algorithm is $\sum_{x \in C_i, y \in C_j} Sim(x, y)/(|C_i||C_j|)$. Similar to the complete-link algorithm, the average-link algorithm has a running time of $O(n^2)$ with $O(n^2)$ space [24]. Both the complete-link and average-link algorithms avoid chaining of clusters, and often produce higher quality clusters than single-link.

The hierarchical centroid-link algorithm differs from the other three hierarchical approaches in that it requires the raw data matrix as input. With the centroid-link criterion, clusters are represented by the mean vectors of the clusters. In the context of gene expression data, the mean vector of a cluster consists of expression levels averaged over all the genes in the cluster under each experiment. The cluster similarity of two clusters is the similarity (or dissimilarity) between the mean vectors of the two clusters. The running time is $O(n^2 p \log n)$ in general.[3]

2.1.5 K-means

K-means is another popular clustering algorithm in gene expression analysis. For example, Tavazoie *et al.* [78] applied k-means to cluster the yeast cell cycle data. K-means [55] is a classic iterative clustering algorithm, in which the number of clusters, k, together with the similarity matrix are inputs to the algorithm. In the k-means clustering algorithm, clusters are represented by *centroids*, which are cluster centers. The goal of k-means is to minimize the sum of distances from each object to its corresponding centroid. In each iteration, each gene is assigned to the centroid (and hence cluster) with the minimum distance (or equivalently maximum similarity). After the gene reassignment, new centroids of the k clusters are computed. The steps of assigning genes to centroids and computing new centroids are repeated until no genes are moved between clusters (and centroids are not changed). K-means was shown to converge for any metric [74].

[3]The hierarchical centroid-link algorithm has a time complexity of $O(n^2)$ when the dimensionality p is fixed and when Euclidean distance is used as the similarity metric [24].

Initialization plays an important role in the k-means algorithm. In one initialization approach, the k initial centroids consist of randomly chosen genes. Another initialization approach is to use the clusters from another clustering algorithm as initial clusters. The advantage of the second approach is that the algorithm becomes deterministic (the algorithm always yields the same clusters). In our implementation, we used the output from the hierarchical average-link algorithm as initial clusters for k-means.

2.1.6 CAST

The *Cluster Affinity Search Technique (CAST)* [10], [9] is a graph-theoretic algorithm developed to cluster gene expression data. In graph-theoretic clustering algorithms, the objects to be clustered (genes in this case) are represented as nodes, and pairwise similarities of genes are represented as weighted edges in a graph. The inputs to CAST include the similarity matrix Sim, and a threshold parameter t (which is a real number between 0 and 1).

CAST is an iterative algorithm in which clusters are constructed one at a time. The current cluster under construction is called C_{open}. The *affinity* of a gene g, $a(g)$, is defined to be the sum of similarity values between g and all the genes in C_{open}, *i.e.*, $a(g) = \sum_{x \in C_{open}} Sim(g, x)$. A gene g is said to have high affinity if $a(g) \geq t|C_{open}|$. Otherwise, g is said to have low affinity. Note that the affinity of a gene depends on the genes that are already in C_{open}. When a new cluster C_{open} is started, the initial affinity is zero because C_{open} is empty. A gene not yet assigned to any clusters and having the maximum average similarity to all unassigned genes is chosen to be the first gene in C_{open}. The algorithm alternates between adding high affinity genes to C_{open}, and removing low affinity genes from C_{open}. C_{open} is *closed* when no more genes can be added to or removed from it. Once a cluster is closed, a new C_{open} is formed. The algorithm iterates until all the genes have been assigned to clusters and the current C_{open} is closed. After the CAST algorithm converges (assuming it does), there is an additional iterative step, in which all clusters are considered at the same time, and genes are moved to the cluster with the highest average similarity.

Correlation coefficient is usually used as the similarity metric for CAST. The iterative

step in CAST usually does not converge if Euclidean distance is used as the similarity metric.

2.2 Data Pre-processing

Data pre-processing plays an important role in clustering gene expression data. Usually, the raw gene expression data set is pre-processed before a clustering algorithm is applied. In gene expression analysis, there are two common types of data pre-processing: filtering and data transformation.

Due to the large number of genes to be clustered, the full gene expression data set is usually *filtered* to reduce the size of the data. Filtering removes genes that do not vary significantly across the experiments. Filtering also facilitates biological interpretation because genes that do not vary significantly across the experiments are usually not of great interest to biologists.

After filtering, gene expression data sets are usually transformed before clustering is applied. The two most popular data transformations include the *logarithm transformation* and *standardization*. In the log transformation, the logarithm function is applied to each expression level in the data. Many statistical techniques assume the data to be normally distributed. The log transformation is known to improve normality of expression data. In standardization (or sometimes called normalization in the literature), the expression vectors are standardized to have mean 0 and standard deviation 1 (by subtracting the mean of each row in the data, and then dividing by the standard deviation of the row). It can be shown that the correlation coefficient and Euclidean distance are equivalent on a standardized data set. Data transformations and normality will be discussed in more detail in Chapter 5.

2.3 Data Sets

In order to evaluate analytical techniques, we used data sets with external criteria (which serve as ideal clusters of the data). Three real gene expression data sets for which external evaluation criteria were available, and three sets of synthetic data were used to assess the quality of clustering results. We use the term *class* to refer to an ideal cluster from the

external criterion. The word *cluster* refers to a cluster obtained by a clustering algorithm.

2.3.1 Gene expression data sets

Ovary data: We used a subset of the ovary data obtained by Schummer *et al.* ([72], [71]). 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 4 different genes. These 4 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.

Yeast cell cycle data: The yeast cell cycle data [20] showed the fluctuation of expression levels of approximately 6000 genes over two cell cycles (17 time points). We used two different subsets of this data with independent external criteria. The first subset (the 5-phase criterion) consists of 384 genes whose expression levels peak at different time points corresponding to the five phases of cell cycle [20]. We expect clustering results to approximate this five class partition. Hence, we used the 384 genes with the 5-phase criterion as one of our data sets. The second subset (the MIPS criterion) consists of 237 genes corresponding to four categories in the MIPS database [60]. 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 [78].

Rat CNS data: The rat CNS data set was obtained by reverse transcription-coupled PCR to study the expression levels of 112 genes during rat central nervous system development over 9 time points [81]. Wen *et al.* [81] classifies the 112 genes into four functional categorizations based on prior biological knowledge. These four functional categories form the four classes in the external criterion in this data set.

2.3.2 Synthetic data sets

Since real expression data sets are expected to be noisy and their clusters may not fully reflect the class information which is derived from information other than gene expression data, 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 three sets of synthetic data, each of which has different properties. By using all three sets of synthetic data, we hope to evaluate the performance of the analytical techniques 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.

In our experiments reported in subsequent chapters, ten replicates are generated from each of the three sets of synthetic data. In each replicate, 235 observations and 24 variables are randomly generated unless otherwise stated.

Mixture of normal distributions based on the ovary data: Visual inspection of the ovary data suggests that the marginal distribution of the expression levels from each experiment is not too far from normal. The expression levels for different clones of the same gene are not identical because the clones represent different portions of the cDNA. Figure 2.1 shows the distribution of the expression levels in a normal tissue from a gene (class) of the ovary data. We found that the distributions of the normal tissue samples are typically closer to the normal distribution than those of tumor tissue samples. A typical example of the distribution of the expression levels in a tumor tissue is shown in Figure 2.2. Even though some of the tumor tissues from some classes (genes) do not closely follow the normal distribution, we generate the replicates using a mixture of multivariate normal distributions in this synthetic data set.

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 2.2). This synthetic data set preserves the mean vector and the covariance matrix between the experiments in each class, but it assumes that the

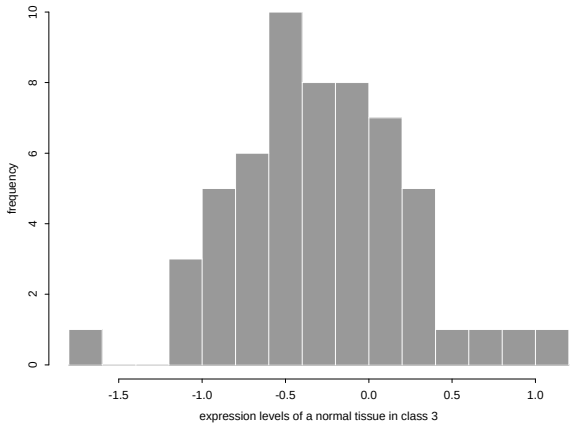

Figure 2.1: Histogram of the distribution of the expression levels in a normal tissue from a gene (class) in the ovary data

underlying distribution of expression levels in each class is multivariate normal.

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.

Cyclic data: This synthetic data set models sinusoidal cyclic behavior of genes over time.[4] 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 $\phi(i,j) = \sin(\frac{2\pi j}{8} - w_k + \epsilon)$. α_i

[4]We would like to thank Dr. Lue Ping Zhao at the Fred Hutchinson Cancer Research Center for suggesting this cyclic model of gene expression data. This model is a simplified form of Dr. Zhao's recent work [84].

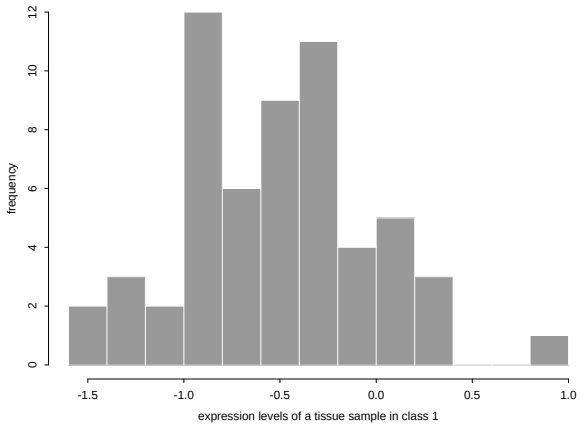

Figure 2.2: Histogram of the distribution of the expression levels in a tumor tissue from another gene (class) in the ovary data

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). There are a total of 10 classes, and k is the class number. The sizes of the different classes are generated according to Zipf's Law [85]. Different classes are represented by different phase shifts w_k, which are chosen according to the uniform distribution in the interval $[0, 2\pi]$. The random variable ϵ, which represents the noise of gene synchronization, is generated according to the standard normal distribution. The parameter λ_j is the amplitude control of condition j, and is simulated according to the normal distribution with mean 3 and standard deviation 0.5. The quantity δ_j, which represents an additive experimental error, is generated from the standard normal distribution. Each observation (row) is standardized to have mean 0 and variance 1.

2.3.3 Summary

The ovary data, yeast cell cycle data and rat CNS data sets were used because the external criteria are available so that we can assess the results from our methodologies. One potential

drawback for these real data is that only a small subset from the full data is used. In addition, selecting a subset of genes that belong to an external criterion may lead to a biased data set. The synthetic data sets complement the drawbacks of real data. The synthetic data are generated with the class information, so we expect the clustering results from the synthetic data to capture more class structure than those from the real data. With the synthetic data, we can also generate larger data sets. For example, in Chapter 5, we generated the mixture of normal synthetic data with 2350 genes. Unlike the ovary data and yeast cell cycle data, the cyclic data contains very small classes (with only a few genes). By using all six real and synthetic data sets, we hope to cover many different properties of real expression data.

2.4 Independent Assessment of Cluster Quality

An underlying theme of this dissertation is evaluation of analytical approaches by comparing clustering results to the corresponding external criteria. In this section, the statistic used to compute the agreement of a clustering result and an external criterion is described.

2.4.1 Adjusted Rand index

Both clustering results and classes in external criteria can be considered as partitions of objects into groups. Thus, comparing a clustering result to an external criterion is equivalent to assessing the agreement of two partitions. The *adjusted Rand index* [39] assesses the degree of agreement between two partitions. Milligan and Cooper [62] recommended the adjusted Rand index as the measure of agreement even when comparing partitions with 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 $1 \leq j \neq j' \leq C$. Suppose that U is our external criterion and V is a clustering result. Let a be the number of pairs of objects that are placed in the same class in U and in the same cluster in V, b be the number of pairs of objects in the same class in U but not in the same cluster in V, c be the number

of pairs of objects in the same cluster in V but not in the same class in U, and d be the number of pairs of objects in different classes and different clusters in both partitions. The quantities a and d can be interpreted as agreements, and b and c as disagreements. The *Rand index* [66] is simply $\frac{a+d}{a+b+c+d}$. The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1.

The problem with the Rand index is that the expected value of the Rand index of two random partitions does not take a constant value (say zero). The adjusted Rand index proposed by Hubert and Arabie [39] assumes the generalized hypergeometric distribution as the model of randomness, *i.e.,* the U and V partitions are picked at random such that the numbers of objects in the classes and clusters are fixed. Let n_{ij} be the number of objects that are in both class u_i and cluster v_j. Let $n_{i.}$ and $n_{.j}$ be the number of objects in class u_i and cluster v_j respectively. The notations are illustrated in a contingency table shown in Table 2.1.

Table 2.1: Notations of a contingency table for comparing two partitions.

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

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

Under the generalized hypergeometric model, it can be shown [39] that:

$$E\left[\sum_{i,j}\binom{n_{ij}}{2}\right] = \left[\sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\right] / \binom{n}{2} \tag{2.1}$$

The expression $a + d$ can be simplified to a linear transformation of $\sum_{i,j}\binom{n_{ij}}{2}$. With

simple algebra, the adjusted Rand index can be simplified to:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}} \tag{2.2}$$

2.4.2 An example

In this section, an example is used to illustrate how the adjusted Rand index is computed. The 10 objects are partitioned into 3 clusters and 3 classes respectively in this example. The contingency table for the example is shown in Table 2.2.

Table 2.2: Contingency table for the example illustrating the adjusted Rand index.

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

a is defined as the number of pairs of objects in the same class in U and same cluster in V, hence a can be written as $\sum_{i,j}\binom{n_{ij}}{2}$. In the example in Table 2.2, $a = \binom{2}{2} + \binom{4}{2} = 7$. b is defined as the number of pairs of objects in the same class in U but not in the same cluster in V. In terms of the notations in Table 2.1, b can be written as $\sum_i \binom{n_{i.}}{2} - \sum_{i,j}\binom{n_{ij}}{2}$. In our example, $b = \binom{2}{2} + \binom{4}{2} + \binom{4}{2} - 7 = 6$. Similarly, c is defined as the number of pairs of objects in the same cluster in V but not in the same class in U, so c can be written as $\sum_j \binom{n_{.j}}{2} - \sum_{i,j}\binom{n_{ij}}{2} = \binom{2}{2} + \binom{3}{2} + \binom{5}{2} - 7 = 7$. d is defined as the number of pairs of objects that are not in the same class in U and not in the same cluster

in V. Since $a + b + c + d = \binom{n}{2}$, $d = \binom{10}{2} - 7 - 6 - 7 = 25$. The Rand index for comparing the two partitions in our example is $\frac{7+25}{45} = 0.711$, while the adjusted Rand index is $\frac{7-14*13/45}{(14+13)/2-14*13/45} = 0.313$. The Rand index is usually higher than the adjusted Rand index. This is because the Rand index lies between zero and one, while the expected value of the adjusted Rand index has value zero and its maximum value is also one. Hence, the adjusted Rand index has a wider range of values.

Chapter 3

COMPARING HEURISTIC-BASED CLUSTERING ALGORITHMS

Many clustering algorithms have been proposed for gene expression data. However, no clustering method has emerged as the method of choice in the gene expression community. Different clustering algorithms can potentially generate different clusters on the same data set. A biologist with a gene expression data set is faced with the problem of choosing an appropriate clustering algorithm for his or her data set. In this chapter, we will describe a data-driven methodology to compare the performance of clustering algorithms.

3.1 Our Approach

Our method for assessing the quality of clustering results is motivated by the jackknife approach [26]. Our idea is to apply a clustering algorithm to all but one experiment in a given data set, and use the left-out experiment to assess the predictive power of the clustering algorithm. We define a scalar quantity called the *figure of merit* (FOM), which is an estimate of the predictive power of a clustering algorithm.

3.1.1 Figure of Merit

Intuitively, a clustering result has possible biological significance if genes in the same cluster tend to have similar expression levels in additional experiments that were not used to form the clusters. We estimate this predictive power by removing one experiment from the data set, clustering genes based on the remaining data, and then measuring the within-cluster similarity of expression values in the left-out experiment. Using our notations from Chapter 2, suppose experiment e is the left-out experiment, and a clustering algorithm is applied to the data under experiments $1, \ldots, (e-1), (e+1), \ldots, p$ to produce k clusters, $C_1, C_2, \ldots, C_k$. Let $D(g, e)$ be the expression level of gene g under experiment e in the raw

data matrix. Let $\mu_{C_i}(e)$ be the average expression level in experiment e of genes in cluster C_i, *i.e.*, $\mu_{C_i}(e) = \sum_{g \in C_i} D(g,e)/|C_i|$. The figure of merit is defined as the sum of squared deviation in the left-out experiment e of the individual gene expression levels relative to their cluster means. Specifically, let $FOM(e,k)$ denote the figure of merit for k clusters with experiment e as validation:

$$FOM(e,k) = \frac{1}{n} \sum_{i=1}^{k} \sum_{g \in C_i} (D(g,e) - \mu_{C_i}(e))^2 \tag{3.1}$$

The figure of merit, $FOM(e,k)$, measures the mean squared error of predicting the expression levels from the average cluster expression level in experiment e. Hence, a relatively *small* figure of merit indicates a clustering algorithm having relatively *high* predictive power.

Each of the p experiments can be used as the validation experiment. The *aggregate figure of merit*, $FOM(k) = \sum_{e=1}^{m} FOM(e,k)$, is an estimate of the total predictive power of the algorithm over all the experiments for k clusters in a data set.

3.1.2 How many clusters are really present?

Ideally, we would like to be able to compare clustering results having different numbers of clusters. Unfortunately, determining the "right" number of clusters in a given data set is a long-standing and very difficult problem [41]. The gap statistic of Tibshirani *et al.* [79] is a recent attempt to estimate the number of clusters in gene expression data sets by comparing within-cluster dispersion to that of a reference null distribution. However, in the absence of a well-grounded statistical model, it seems impossible to determine the "right" number of clusters, or even to define the concept.[1] Therefore, we compare the figure of merit of different clustering algorithms over a range of different numbers of clusters.

Clustering algorithms typically have parameters that directly (for example, k-means) or indirectly (for example, CAST) determine the number of clusters. To compare the figures of merit of clusters produced by two different algorithms, we adjust the parameters so that the number of clusters is the same in both cases.

---

[1] We propose a probability framework for clustering gene expression data in Chapter 5.

3.1.3 Adjusted Figure of Merit

Since the figure of merit, $FOM(e,k)$, is defined as the sum of within-cluster variance over all the k clusters, $FOM(e,k)$ decreases as the number of clusters increases. Hence, we define the *adjusted figure of merit*, which is the figure of merit divided by a factor that compensates for the statistical bias from the numbers of clusters.

Our approach is to derive the adjustment factor that compensates for the statistical bias by assuming an idealized model. Suppose that the n genes fall into c true classes, with the ith class containing $\alpha_i n$ genes, where $0 < \alpha_i < 1$ and $\sum_{i=1}^{c} \alpha_i = 1$. Further assume that the expression levels of genes in class i under experiment e are independent normally distributed random variables with mean $\mu_{i,e}$ and variance $\sigma_{i,e}^2$.

Suppose we apply a clustering algorithm to the n genes to obtain k clusters, where $k \geq c$. We assume that the clustering algorithm is perfect, in the sense that each cluster contains genes from only one class. Assume there are $\alpha_i k$ clusters containing class i genes. (This assumption is valid if the clustering algorithm favors equal-sized clusters. However, the analysis is otherwise independent of the sizes of the clusters within each class.)

Theorem 1 *With the above assumptions, the expected aggregate FOM, $E[FOM(k)]$, is $\frac{n-k}{n}\bar{\sigma}$, where $\bar{\sigma}$ is a weighted average of the $\sigma_{i,e}$, independent of k. Specifically, $\bar{\sigma} = \sum_{e=1}^{p} \sum_{i=1}^{c} \alpha_i \sigma_{i,e}^2$.*

Proof Outline: Suppose the measured expression levels of the $\alpha_i n$ genes in true class i under experiment e are $x_{1,e}, \ldots, x_{\alpha_i n,e}$. Let $\bar{x}_e = \sum_{i=1}^{\alpha_i n} x_{i,e}/\alpha_i n$. Then the expected value of $\sum_{i=1}^{\alpha_i n}(x_{i,e} - \bar{x}_e)^2$ is $(\alpha_i n - 1)\sigma_{i,e}^2$. Subdividing this cluster into $\alpha_i k$ smaller nonempty sub-clusters would reduce these genes' expected contribution to the $FOM(e,k)$ to $(\alpha_i n - \alpha_i k)\sigma_{i,e}^2$. Hence, $E[FOM(k)] = \sum_{e=1}^{p} E[FOM(e,k)] = \sum_{e=1}^{p} \frac{1}{n} \sum_{i=1}^{c} (\alpha_i n - \alpha_i k)\sigma_{i,e}^2 = \frac{n-k}{n} \sum_{e=1}^{p} \sum_{i=1}^{c} \alpha_i \sigma_{i,e}^2$. □

If the assumptions in Theorem 1 are satisfied and round-off errors ($\alpha_i k$ is assumed to be an integer) are ignored, the rate of decline of $FOM(k)$ as k, the number of clusters, increases should be $\frac{n-k}{n}$. The *adjusted figure of merit* of k clusters is defined to be $FOM(e,k)/\frac{n-k}{n}$.

3.1.4 Related Work

Our approach has some similarity to *leave-one-out cross validation* in machine learning. In leave-one-out cross validation, the objective is to estimate the accuracy of a *classifier*, an algorithm that maps an unlabeled instance to a label, by *supervised learning* [47]. The labels of the objects to be classified are assumed to be known. The idea is to hide the label of each object in turn, and to estimate the label of the object using a classifier. This is in contrast to our approach in which we do *not* assume any prior information of the genes to evaluate the quality of clustering results. Instead, we define figures of merit, which are estimators of the predictive power of clustering algorithms, to assess the quality of clustering results.

3.2 Experimental Details

3.2.1 Clustering Algorithms

We implemented two partitional clustering algorithms (CAST and k-means), and four hierarchical clustering algorithms (single-link, average-link, centroid-link and complete-link). These clustering algorithms are described in Section 2.1.3 in Chapter 2. We also investigated another popular clustering algorithm called self-organizing map (SOM), and found that the performance of SOM is dependent on many different parameters. Hence, the results of SOM are not reported in this chapter. Please refer to Appendix A for more details of SOM. We used the correlation coefficient as the similarity metric in all our experiments. Moreover, we implemented the *random clustering algorithm* as a benchmark for evaluating the performance of other clustering algorithms. A random clustering with k clusters is obtained by placing k randomly selected genes into separate clusters, then assigning the remaining genes to the k clusters uniformly at random. An algorithm whose figure of merit is little better than that of a random clustering is probably producing poor clusters.

3.2.2 Data Sets

We applied our FOM methodology to the real and synthetic data sets for which external criteria are available. The data sets are described in Section 2.3 in Chapter 2.

3.3 Results and Discussion

In this section, we showed the results of applying our FOM methodology to various real and synthetic data sets using different clustering algorithms. In our experiments, the random clustering algorithm was repeated 1000 times, and the average FOM's over these 1000 random runs are shown in the following results.

3.3.1 Real data sets

The Ovary Data Set:

Figure 3.1 shows the adjusted FOM's on the ovary data set from 1 to 30 clusters. The adjusted FOM's of the random algorithm are almost constant over different numbers of clusters. This shows that the adjustment factor $\frac{n-k}{n}$ derived from the idealized model in Section 3.1.3 compensates for the statistical effect of increasing the number of clusters despite the fact that the data and the random algorithm probably violate key assumptions in the analysis. However, the adjustment factor $\frac{n-k}{n}$ is not perfect as the adjusted FOM's of all other clustering algorithms decrease as the number of clusters, k, increases. From Figure 3.1, the hierarchical single-link algorithm produces FOM's that are only slightly better than the random algorithm. This observation is consistent with the general belief that single-link is less desirable than the other three hierarchical approaches. This is because single-link can potentially cause chaining of clusters (described in Section 2.1.3). Figure 3.1 also shows that centroid-link produces lower FOM's than single-link, but higher FOM's than the other clustering algorithms. The complete-link, CAST and k-means algorithms achieve the lowest FOM's on this data. Both the complete-link and CAST algorithms show a steep decline of FOM's up to around 4 to 6 clusters. Since the 235 clones in this data set correspond to 4 genes, our result gives a reasonable hint to the correct number of clusters.

The Yeast Cell Cycle Data Set (5-phase criterion):

Figure 3.2 shows the adjusted FOM's of seven clustering algorithms for 1 to 30 clusters on the yeast cell cycle data set with the 5-phase criterion. The random algorithm again produces near-constant adjusted FOM's over different numbers of clusters. The hierarchical single-link algorithm again produces only slightly better clusters than the random algorithm.

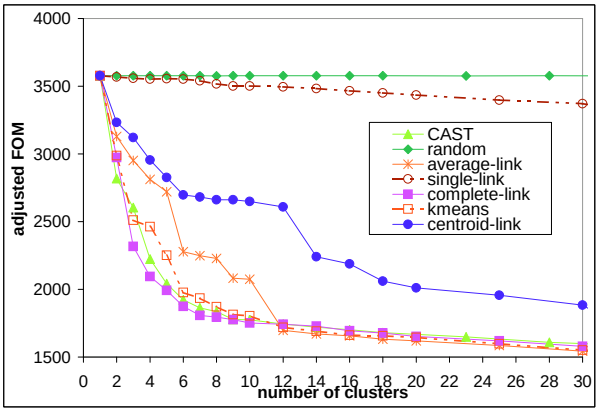

Figure 3.1: Adjusted FOM's on the ovary data set.

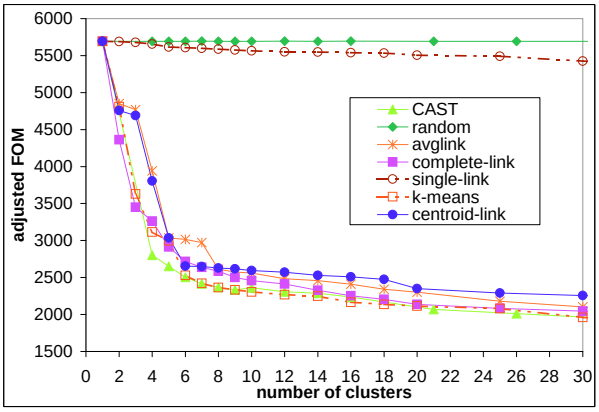

Figure 3.2: Adjusted FOM's on the yeast cell cycle data.

The hierarchical centroid-link and average-link produce clustering results with comparable FOM's. The CAST, hierarchical complete-link and k-means algorithms produce the lowest FOM's. The adjusted FOM's of CAST, k-means and complete-link show a steep decline until four to six clusters, which is consistent with the five classes corresponding to the five phases of cell cycle identified by Cho *et al.* [20]. In particular, CAST produces the lowest adjusted FOM's at four and five clusters.

The Rat CNS Data Set:

Figure 3.3 shows the adjusted FOM's for 1 to 30 clusters on the rat CNS data. The results are similar to those on the ovary data and the yeast cell cycle data, with the random algorithm producing almost constant adjusted FOM's, and hierarchical complete-link,

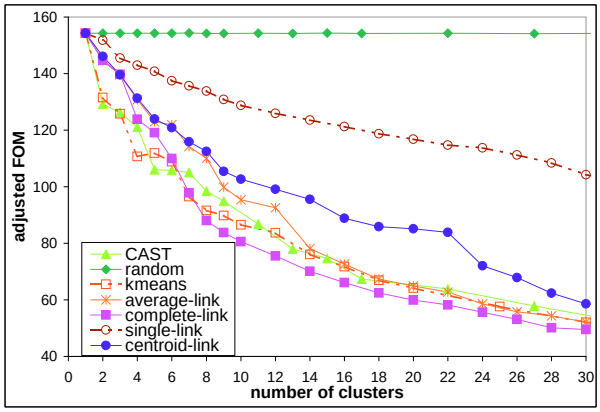

Figure 3.3: Adjusted FOM's on the rat CNS data.

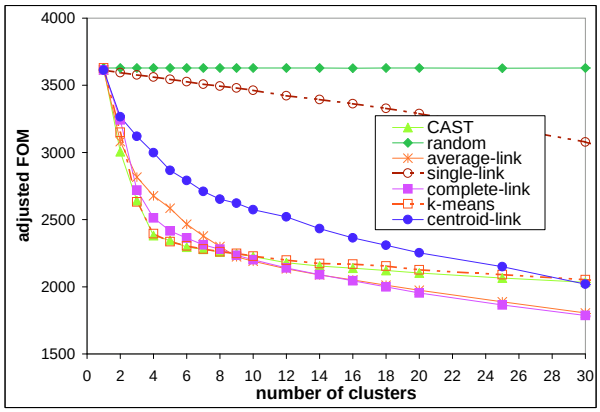

Figure 3.4: Average adjusted FOM's on the mixture of normal distributions synthetic data.

CAST and k-means producing the lowest FOM's. However, hierarchical single-link produces significantly higher quality clusters than the random algorithm.

3.3.2 Synthetic data sets

Mixture of normal distributions based on the ovary data:

Figure 3.4 shows the average adjusted FOM's over ten replicates of the mixture of normal distributions based on the ovary synthetic data. The general trends are very similar to the real data sets, with CAST, k-means and complete-link producing the highest quality clusters. Moreover, the decline of adjusted FOM's starts to be less steep around 4 clusters, which is the number of classes for this data.

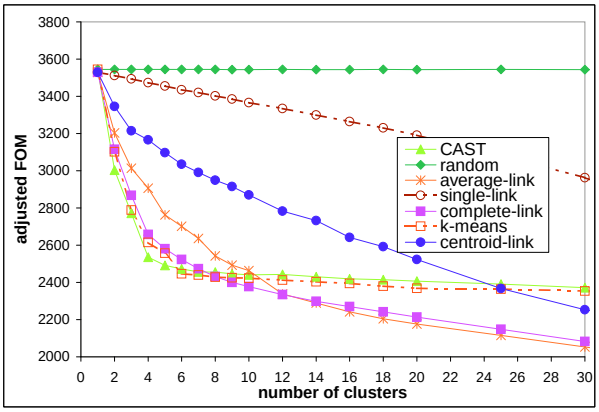

Figure 3.5: Average adjusted FOM's on the randomly resampled synthetic data.

Randomly resampled ovary data:

Figure 3.5 shows the average adjusted FOM's over ten replicates of the randomly resampled ovary synthetic data. The results are very similar to that on the mixture of normal distributions based on the ovary data. There are four classes in this data set, and we again observed that the adjusted FOM's decline less steeply around 4 clusters. In particular, CAST produces the lowest FOM's at 4 and 5 clusters.

Cyclic data:

Figure 3.6 shows the average adjusted FOM's over ten replicates of the cyclic data. This data set shows a very different picture than the other synthetic data sets: all the clustering algorithms except single-link produce very similar adjusted FOM's, and single-link produces considerably lower FOM's than random. The cyclic data set represents the most "difficult" synthetic data because it contains very small classes. Moreover, the decline in adjusted FOM's starts to be less steep around 4 to 6 clusters, but the number of classes in this data is 10.

3.4 Validation of FOM methodology

Validating clustering results is a well-studied problem in statistics. Jain and Dubes [40] classified cluster validation procedures into two main categories: external and internal criterion analysis. *External criterion* analysis validates a clustering result by comparing it

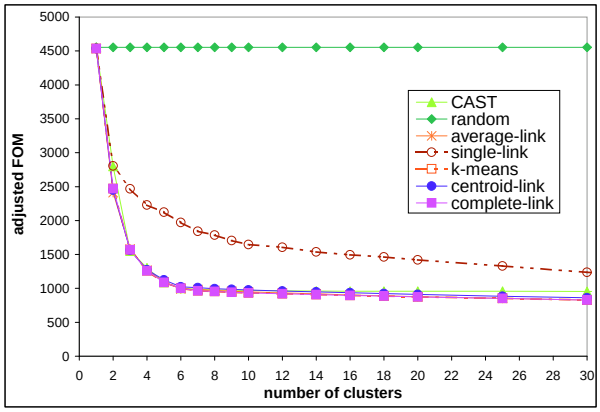

Figure 3.6: Average adjusted FOM's on the cyclic data.

to a given "gold standard" which is another partition of the objects. The gold standard must be obtained by an independent process based on information other than the given data. There are many statistical measures that assess the agreement between an external criterion and a clustering result, for example, the adjusted Rand index [39] described in Chapter 2. However, reliable external criteria are rarely available when analyzing gene expression data. *Internal criterion* analysis uses information from within the given data set to represent the goodness of fit between the input data set and the clustering results. For example, homogeneity of genes within clusters and separation between clusters are possible measures of goodness of fit [75].

For validation of clustering results, external criterion analysis has the strong benefit of providing an independent, hopefully unbiased assessment of cluster quality. On the other hand, external criterion analysis has the strong disadvantage that an external gold standard is rarely available. Internal criterion analysis avoids the need for such a standard, but has the alternative problem that clusters are validated using the same information from which clusters are derived. Different clustering algorithms optimize different objective functions or criteria. Assessing the goodness of fit between the input data set and the resulting clusters is equivalent to evaluating the clusters under a different objective function. Our FOM approach compromises these two extremes: *no* external standard is required, and the clustering results are evaluated based on homogeneity of the *hidden* data that are not available to the clustering algorithms.

Since external criteria are available for all of the real and synthetic data sets we used, we are able to evaluate the clustering results using the external criteria, and to compare the conclusions to those obtained from the FOM approach in Section 3.4.1. Moreover, we also compared our FOM approach to other cluster validation methodologies that do not require external criteria in Section 3.4.2.

3.4.1 External Validation

The goal of this section is to contrast the results from our FOM approach to those from the external criteria. In our FOM approach, the clustering results are obtained by leaving out one experiment, and using the left-out experiment to evaluate the clustering result. The aggregate FOM for k clusters, $FOM(k)$, is equivalent to the average FOM over all the possible left-out experiments. In order to validate the FOM approach, we compared the clustering results from leaving out one experiment to the external criterion in the same manner. Specifically, we applied clustering algorithms to experiments $1, \ldots (e-1), (e+1), \ldots, p$ (where $e = 1 \ldots p$) to produce k clusters, and compared the clustering results to the external criterion using the adjusted Rand index. Then, we computed the average adjusted Rand index by averaging over all the experiments e for k clusters. In the following results, the average adjusted Rand indices are plotted against the number of clusters k on both real and synthetic data sets.

The Ovary Data Set:

Figure 3.7a shows the average adjusted Rand indices for 1 to 30 clusters on the ovary data. The results from the random algorithm are not shown because the average adjusted Rand indices are consistently zero. Similar to the results from the FOM approach (Figure 3.1), single-link produces clustering results with very low adjusted Rand indices, and hence poor clustering results. CAST, complete-link and k-means produce relatively high adjusted Rand indices. The number of classes in this data set is four, and the average adjusted Rand indices reach the maximum around 4 to 6 clusters (the same range as the decline of FOM's starting to be less steep).

Figure 3.7b plots the adjusted FOM against the average adjusted Rand indices at fixed

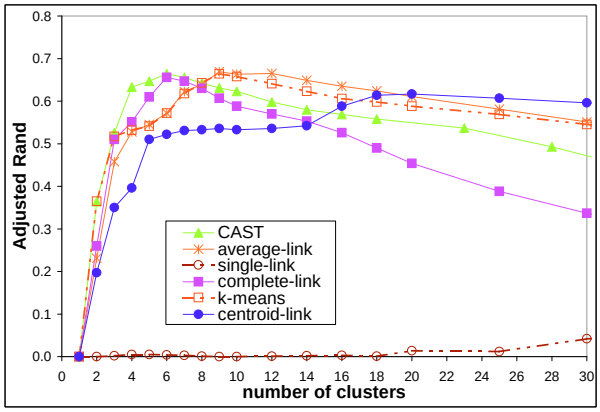

Figure 3.7a: Average adjusted Rand indices over different numbers of clusters on the ovary data.

numbers of clusters, *i.e.*, a point in the graph with average adjusted Rand index $adjR$ and adjusted FOM $adjFOM$ means that the average adjusted Rand index is $adjR$ and the adjusted FOM is $adjFOM$ for a clustering result with k clusters. In Figure 3.7b, the number of clusters increases for the points from the upper left corner of the graph down to the horizontal U-turn in the lower right corner, and then from the U-turn down to the lower left corner of the graph. This means that clustering results with low FOM's tend to have high adjusted Rand indices for small numbers of clusters. The horizontal U-turn indicates the numbers of clusters at which the decline of FOM becomes less steep and the average adjusted Rand index reaches the maximum. The trend from the U-turn to the lower left corner indicates that the adjusted FOM's continue to decrease while the average adjusted Rand indices start to decrease as the number of clusters is increased.

The Yeast Cell Cycle Set (5-phase criterion):

Figure 3.8a shows the average adjusted Rand indices for 1 to 30 clusters on the yeast cell cycle data with the 5-phase criterion. Again, the results are consistent with the results from the FOM approach, with CAST and k-means producing the highest quality of clusters between 3 to 5 clusters. Figure 3.8b plots the adjusted FOM against the average adjusted Rand indices at fixed numbers of clusters. The downward trend corresponds to small numbers of clusters, while the upward trend corresponds to large numbers of clusters.

The Rat CNS Set:

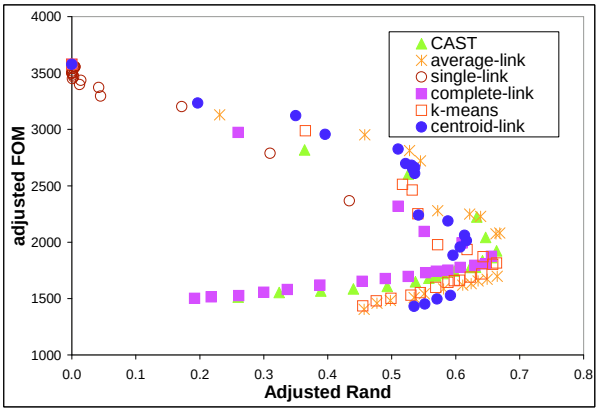

Figure 3.7b: Adjusted FOM against average adjusted Rand indices on the ovary data.

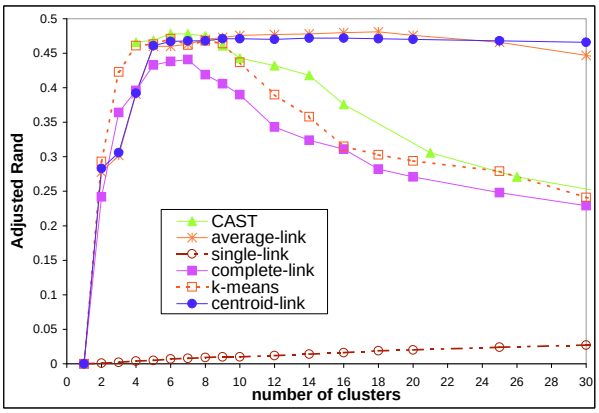

Figure 3.8a: Adjusted Rand indices over different numbers of clusters on the yeast cell cycle data.

Figure 3.9 shows the average adjusted Rand indices from 1 to 30 clusters on the rat CNS data. The results on this data are not as consistent with the FOM approach as the ovary and yeast cell cycle data. This is probably due to the fact that the external criterion is not as reliable as the ovary and yeast cell cycle data: the maximum adjusted Rand index is only about 0.20, while the maximum average adjusted Rand indices are close to 0.50 and 0.70 for the yeast cell cycle and ovary data respectively. The low average adjusted Rand index on this data shows that the clustering results cannot capture the external criterion, and hence the external criterion is not an objective way to validate the clustering results for this data.

Mixture of normal distributions based on the ovary data:

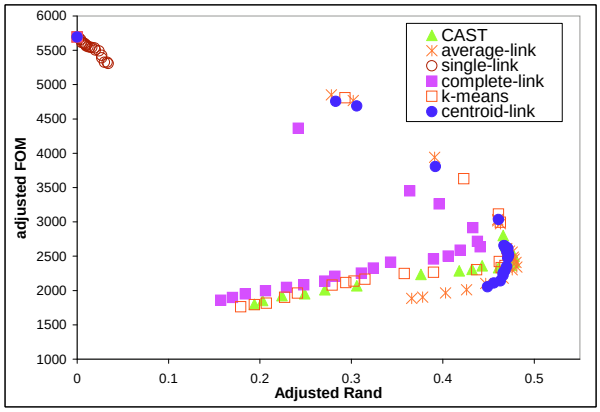

Figure 3.8b: Adjusted FOM against adjusted Rand indices on the yeast cell cycle data.

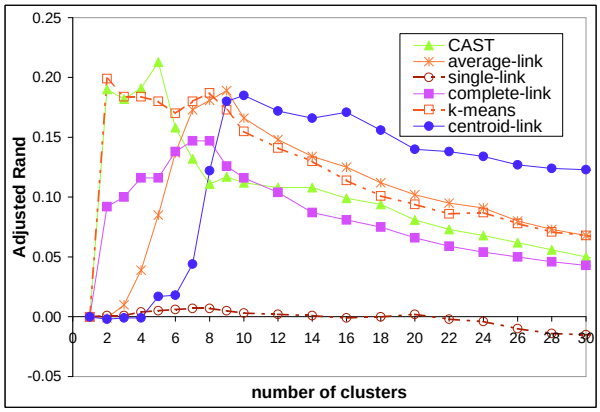

Figure 3.9: Adjusted Rand indices over different numbers of clusters on the rat cns data.

Figure 3.10a shows the average adjusted Rand indices over the ten replicates from 1 to 30 clusters on the mixture of normal distributions data. The results are again consistent with the results from the FOM approach, with CAST and k-means producing relatively high quality clusters. The number of classes in this synthetic data set is four. CAST and k-means produce the maximum average adjusted Rand indices at four clusters. Figure 3.10b plots the average adjusted FOM against the average adjusted Rand indices at fixed numbers of clusters. There is a downward trend for small numbers of clusters, showing that clustering results with low FOM's tend to have high agreement to the external criterion. The upward trend for large numbers of clusters reflects that the adjustment factor $\frac{n-k}{n}$ is not perfect, *i.e.*, the adjusted FOM's continue to decrease as the number of clusters is increased past

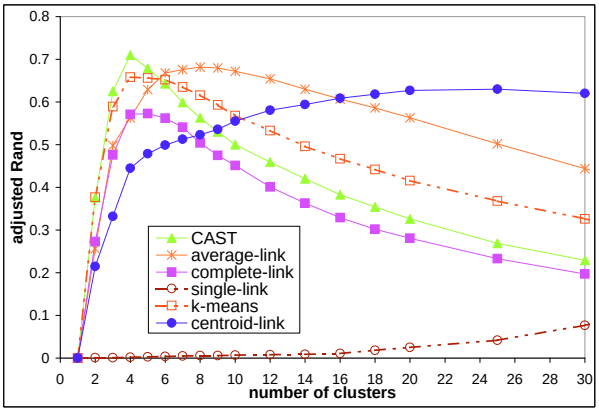

Figure 3.10a: Average adjusted Rand indices over different numbers of clusters on the mixture of normal data.

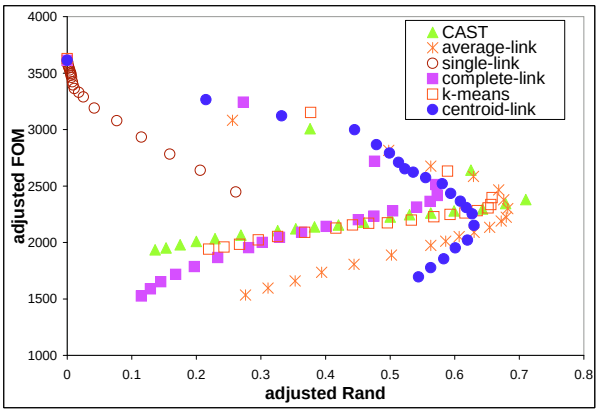

Figure 3.10b: Average adjusted FOM against average adjusted Rand indices on the mixture of normal data.

the true number of classes.

Randomly resampled ovary data:

Figure 3.11 again shows that the FOM approach generates conclusions that are highly consistent with the degree of agreement to the external criterion. The only exception is complete-link, which produces clustering results with comparable FOM's to CAST and k-means, but with lower adjusted Rand indices than CAST and k-means. The average adjusted Rand index again peaks at four, which is the number of classes for this data, for CAST and complete-link. The plot of adjusted FOM against adjusted Rand index is highly similar to that for the mixture of normal distributions data, and is not shown here.

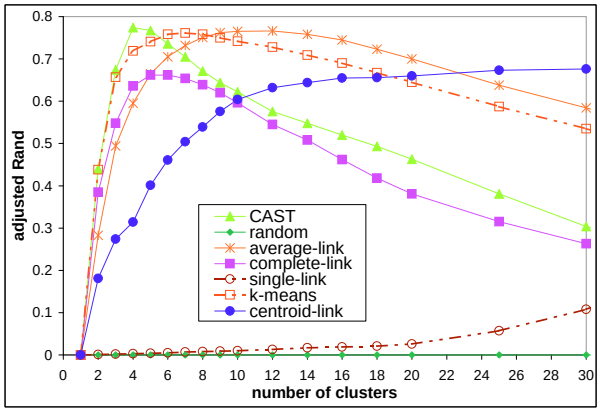

Figure 3.11: Average adjusted Rand indices over different numbers of clusters on the randomly resampled ovary data.

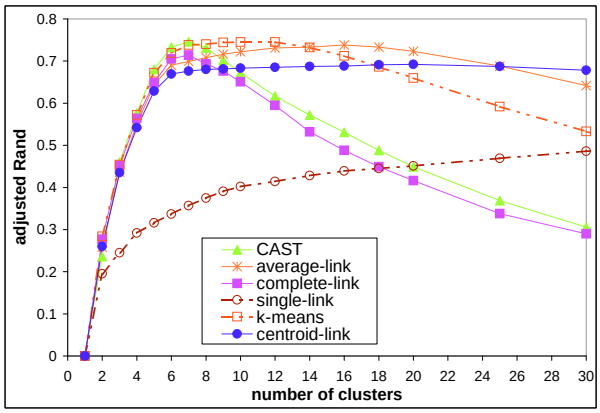

Figure 3.12a: Average adjusted Rand indices over different numbers of clusters on the cyclic data.

Cyclic data:

Figure 3.12a shows that all clustering algorithms (except single-link) produce comparable average adjusted Rand indices until 5 to 6 clusters, which is consistent with the results from the FOM approach. The discrepancies again lie in the number of clusters after the peak of the adjusted Rand index is reached. This is due to the imperfect adjustment factor. Figure 3.12b again shows the same results in a different graph. In general, the FOM decreases as the adjusted rand index increases until around the number of clusters for which the adjusted Rand index peaks.

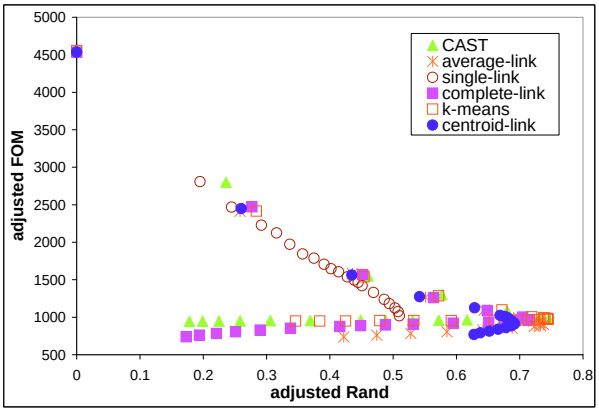

Figure 3.12b: Average adjusted FOM against average adjusted Rand indices on the cyclic data.

3.4.2 Internal Validation

In addition to comparing the results from the FOM approach to those from using external criteria, we also compared the results from the FOM method to other internal validation approaches (in which no external criterion is necessary). The internal validation approaches include the homogeneity and separation criteria by Shamir and Sharan [75], and the silhouette approach by Rousseeuw [68].

Homogeneity and separation

Since objects in the same cluster are expected to be more similar to each other than objects in different groups and objects in different clusters are expected to be dissimilar, homogeneity of objects in the same cluster and separation between different clusters are intuitive measures of cluster quality [75]. Let D_g be the vector of expression levels of gene g under all p experiments, where $g = 1 \ldots n$. Let $Centroid_{C_i}$ be the cluster center of cluster C_i, where $i = 1 \ldots k$. Let $Sim(u, v)$ be the similarity of vector u and vector v. Since we use correlation coefficient as the similarity metric, $Sim(u, v)$ is equivalent to the correlation coefficient between vectors u and v. Homogeneity, H_{avg}, is defined as the average similarity between objects and their cluster centers, *i.e.*,

$$H_{avg} = \frac{1}{n} \sum_{i=1}^{k} \sum_{g \in C_i} Sim(g, Centroid_{C_i}) \tag{3.2}$$

Separation, S_{avg}, is defined as the weighted average similarity between cluster centers, *i.e.*,

$$S_{avg} = \frac{1}{\sum_{i \neq j} |C_i||C_j|} \sum_{i \neq j} |C_i||C_j| Sim(Centroid_{C_i}, Centroid_{C_j}) \tag{3.3}$$

A high homogeneity indicates objects in clusters are similar to each other. A low separation means that different cluster centers have low similarity. Suppose $\mathcal{C}_1$ and $\mathcal{C}_2$ are two clustering results. $\mathcal{C}_1$ is said to be a better clustering result than $\mathcal{C}_2$ if the homogeneity of $\mathcal{C}_1$ is higher than that of $\mathcal{C}_2$ and the separation of $\mathcal{C}_1$ is lower than that of $\mathcal{C}_2$.

Silhouette

Silhouettes can be used to evaluate the quality of a clustering result. Silhouettes are defined for each object and are based on the ratio between the distances of an object to its own cluster and to its neighbor cluster [68]. Suppose gene g has been assigned to cluster A. Let $Dist(u, v)$ be the dissimilarity between object u and object v. Let $a(g)$ denote the average dissimilarity of g to all other objects in cluster A, *i.e.*, $a(g) = \sum_{h \in A, h \neq g} Dist(g, h)/(|A| - 1)$. Suppose C is another cluster that is different from cluster A, and define $d(g, C)$ to be the average dissimilarity of gene g to all objects in cluster C, *i.e.*, $d(g, C) = \sum_{h \in C} Dist(g, h)/|C|$. The average dissimilarities of gene g to all clusters $C \neq A$ are computed, and let $b(g)$ be the minimum dissimilarity, *i.e.*, $b(g) = \min_{C \neq A} d(g, C)$. The cluster B for which the minimum $b(g)$ is attained is called the *neighbor* of object g. In other words, cluster B is the second-best choice for object g other than cluster A. The silhouette $s(g)$ for object g is defined as

$$s(g) = \frac{b(g) - a(g)}{max\{a(g), b(g)\}} \tag{3.4}$$

The value of $s(g)$ lies in the range of -1 and 1. When $s(g)$ is close to 1, the "within" dissimilarity $a(g)$ is much smaller than the smallest "between" dissimilarity $b(g)$. Hence, object g lies well within its cluster. On the other hand, when $s(g)$ is close to -1, the "within" dissimilarity $a(g)$ is much larger than the smallest "between" dissimilarity $b(g)$. Hence, object g should be assigned to another cluster. When $s(g)$ is close to 0, $a(g)$ and $b(g)$ are approximately equal, and hence it is not clear whether object g should be assigned to cluster A or B. The definition of the silhouette requires at least two clusters. When a cluster

contains only a single object, the silhouette for that object is defined to be zero. Since we use correlation coefficient (which is a similarity measure instead of a dissimilarity measure) as the similarity metric, we adopt the transformation of $Dist(u, v) = (1 - Sim(u, v))/2$ (where u, v are vectors) to convert the similarity values into dissimilarity values for the computation of silhouettes.

Silhouettes can be used to visually display clustering results. The objects in each cluster can be displayed in decreasing order of the silhouette values such that a cluster with many objects with high silhouette values is a pronounced cluster. In order to summarize the silhouette values in a data set with k clusters, the *average silhouette width*, $\bar{s}(k)$, is defined to be the average silhouette value over all the objects in the data, *i.e.*, $\bar{s}(k) = \sum_{g=1}^{n} s(g)/n$.

Silhouettes can also be used to estimate the number of clusters in a given data set. The *silhouette coefficient*, SC, is defined to be $\max_k \bar{s}(k)$, where $k = 2, 3, \ldots, (n-1)$. The silhouette coefficient is a measure of the amount of clustering structure discovered by the clustering algorithm.

Results

The results of comparing the FOM approach to other cluster validation approaches with number of clusters equal to the number of classes are summarized in Table 3.1a and Table 3.1b for real and synthetic data sets respectively. For each data set, the first two rows show the average adjusted Rand indices from leaving out each experiment in turn and the adjusted aggregate FOM at the number of classes as described in Section 3.4.1. The last three rows for each data set show the adjusted Rand index, average silhouette width, homogeneity and separation values computed from clustering results obtained using the entire data. The adjusted Rand index shows the agreement of clustering results to the external criterion, and hence is an external validation approach. The average silhouette width, homogeneity and separation values do not require any external criteria, and hence are internal validation approaches as described in Section 3.4.2. The results of six clustering algorithms (CAST, k-means, and four hierarchical clustering approaches), and from the true classes are shown. The clustering approach favored by each validation approach is shown in bold. A

low FOM corresponds to a high quality clustering result, while a high adjusted Rand index and a high average silhouette width corresponds to a high quality clustering result. Hence, the maximum adjusted Rand index and silhouette width over the six clustering algorithms for each data set are shown in bold, and the minimum FOM for each data set is shown in bold. Similarly, a clustering result with a high homogeneity and low separation values are also shown in bold such that there is no other clustering result with a strictly higher homogeneity value and a strictly lower separation value.

One observation from Table 3.1a and Table 3.1b is that the average adjusted Rand indices from leaving out each experiment in turn, Rand(a), are only slightly lower than the adjusted Rand indices using all the experiments, Rand(b). This shows that leaving out one experiment in the FOM approach does not have a significant effect on the quality of clustering results.

On both real and synthetic data sets, the results favored by the FOM approach agree with the results favored by the adjusted Rand indices in general. On the yeast cell cycle data, the FOM approach and the average adjusted Rand index (Rand(a)) favor the same clustering algorithm, CAST, at five clusters. On both the ovary and rat data, the FOM approach selects the clustering algorithm with the second best average adjusted Rand indices. On the synthetic data, the FOM approach agrees with the average adjusted Rand indices on both the mixture of normal and randomly resampled data. In general, the silhouette approach does not show as much agreement to the external validation (the adjusted Rand index computed using the entire data) as the FOM approach. For example, the silhouette approach selects the clustering result with the second highest adjusted Rand index on both the mixture of normal and randomly resampled data.

The homogeneity and separation criteria usually have problems selecting one single clustering approach. A clustering result A is said to be better than clustering result B only if the homogeneity value of A is higher than that of B *and* the separation value of A is lower than that of B. For example, the homogeneity and separation approach cannot decide whether the CAST or complete-link clustering result is better at four clusters on the ovary data because the homogeneity value from CAST is higher ($0.815 > 0.798$) but the separation from CAST is also higher ($0.340 > 0.319$). In some cases, the homogeneity and

Table 3.1a: Comparing the FOM approach to other validation methods at the number of classes on real expressoin data. For each data set, the first two rows correspond to the average adjusted Rand indices, Rand(a), and the adjusted aggregate FOM described in Section 3.4.1. The third row, Rand(b), corresponds to adjusted Rand indices computed using the entire data. The silhouette, homogeneity and separation values are described in Section 3.4.2.

data	*validation method*	*clustering algorithms* CAST	k-means	single-link	average-link	complete-link	centroid-link	*true classes*
ovary	Rand(a)	**0.633**	0.532	0.006	0.528	0.551	0.396	
	FOM	2223.073	2462.549	3553.392	2812.751	**2094.417**	2955.246	2272.176
	Rand(b)	**0.664**	0.543	0.006	0.673	0.556	0.517	
	silhouette	**0.438**	**0.438**	-0.014	0.429	**0.438**	0.393	0.294
	homo, sep	**0.81, 0.340**	0.795, 0.321	0.595, 0.334	0.791, 0.324	**0.798, 0.319**	0.784, 0.328	0.744, 0.416
cell cycle	Rand(a)	**0.466**	0.461	0.004	0.391	0.396	0.392	
	FOM	**2652.197**	2995.957	5614.924	3034.270	2913.545	3035.135	3011.214
	Rand(b)	0.476	0.466	0.008	0.457	**0.483**	0.476	
	silhouette	**0.515**	0.512	-0.167	0.483	0.465	0.252	0.227
	homo, sep	**0.795, -0.222**	**0.796, -0.216**	**0.303, -0.477**	0.786, -0.215	0.777, -0.210	0.789, -0.204	0.714, -0.109
rat	Rand(a)	**0.191**	0.184	0.004	0.039	0.016	-0.001	
	FOM	121.027	**110.709**	142.956	130.794	123.875	131.301	113.100
	Rand(b)	0.199	**0.212**	0.006	-0.003	0.164	-0.003	
	silhouette	**0.366**	0.334	0.012	0.306	0.291		-0.041
	homo, sep	**0.85, 0.592**	0.832, 0.614	0.741, 0.390	**0.771, 0.313**	0.837, 0.62	**0.771, 0.313**	0.763, 0.854

Table 3.1b: Comparing the FOM approach to other validation methods at the number of classes on synthetic data. For each data set, the first two rows correspond to the average adjusted Rand indices, Rand(a), and the adjusted aggregate FOM described in Section 3.4.1. The third row, Rand(b), corresponds to adjusted Rand indices computed using the entire data. The silhouette, homogeneity and separation values are described in Section 3.4.2.

data	*validation method*	*clustering algorithms*						*true classes*
		CAST	k-means	single-link	average-link	complete-link	centroid-link	
normal	Rand(a)	**0.710**	0.658	0.002	0.563	0.571	0.445	
	FOM	**238.695**	2395.881	3561.287	2676.702	2512.550	2937.975	2291.520
	Rand(b)	**0.709**	0.672	0.001	0.551	0.598	0.467	
	silhouette	0.326	**0.328**	-0.117	0.289	0.274	0.261	0.280
	homo, sep	**0.787, 0.395**	**0.787, 0.390**	0.608, 0.482	**0.761, 0.364**	0.775, 0.421	**0.746, 0.339**	0.771, 0.410
re-	Rand(a)	**0.774**	0.719	0.003	0.595	0.636	0.314	
sampled	FOM	**2535.275**	2614.539	3473.252	2906.692	2657.725	316.034	2222.601
	Rand(b)	**0.782**	0.735	0.003	0.603	0.658	0.282	
	silhouette	0.343	**0.345**	-0.050	0.314	0.289	0.179	0.316
	homo, sep	**0.784, 0.408**	**0.779, 0.401**	0.602, 0.403	**0.756, 0.384**	0.773, 0.433	**0.693, 0.329**	0.777, 0.413
cyclic	Rand(a)	0.672	**0.745**	0.402	0.722	0.651	0.683	
	FOM	969.521	975.566	1647.327	948.309	**933.980**	975.674	903.382
	Rand(b)	0.664	**0.768**	0.422	0.727	0.644	0.683	
	silhouette	0.287	0.345	0.058	**0.348**	0.253	0.326	0.218
	homo, sep	**0.920, 0.036**	**0.916, -0.032**	**0.853, -0.368**	**0.914, -0.069**	0.918, 0.036	**0.909, -0.115**	0.917, 0.017

separation approach cannot even decide that single-link is producing much lower quality clusters. For example, on the yeast cell cycle and cyclic synthetic data, single-link produces much lower adjusted Rand indices and average silhouette widths and much higher FOM's than other clustering algorithms, but the homogeneity and separation approach show that the single-link clustering results have low homogeneity values and low separation values.

As expected, the cyclic synthetic data set is the most difficult data for all internal validation approaches. The FOM approach selects the complete-link result, which has the second lowest average adjusted Rand indices. The homogeneity and separation approach cannot decide between five out of six clustering results on this data.

It is interesting to note that the adjusted aggregate FOM's for the true classes are higher than the best FOM's achieved by a clustering algorithm on all three real data sets. However, the adjusted aggregate FOM's for the true classes are lower than the best FOM's achieved by a clustering algorithm on all three synthetic data sets. This is no surprise: we expect the external criteria on synthetic data to fit the data relatively well since the data are generated using the external criteria. On the other hand, the average silhouette widths for the true classes are lower than the average silhouette widths achieved by the best clustering algorithms on all real and synthetic data. Since the silhouette definition favors data points that are assigned to the cluster with the highest average similarity, the silhouette widths may be biased against the potentially overlapping true classes and favor the disjoint clustering results. The average adjusted Rand indices, Rand(a), and adjusted Rand indices, Rand(b), for the true classes are not shown in Table 3.1a and Table 3.1b because the values are 1 by definition.

Table 3.2 shows the number of clusters, k, that maximizes the average silhouette width, $\bar{s}(k)$, for each clustering algorithm on real and synthetic data sets. The clustering algorithms favored by the average silhouette width are shown in bold. For example, the silhouette approach selects CAST at 6 clusters on the ovary data. On real data, the silhouette approach tends to choose the number of clusters that are around the neighborhood of the true number of clusters. A graph plotting the average silhouette width against the number of clusters for the ovary data is shown in Figure 3.13. The results from the silhouette approach are very similar to those from the FOM approach: single-link produces much lower quality clusters

while k-means, CAST and complete-link produce the highest quality clusters. However, the picture is quite different on the synthetic data: the silhouette approach tends to favor very low number of clusters. Figure 3.14 shows that the average silhouette width tends to decrease when the number of clusters is increased on the mixture of normal distributions data. The results are similar on the randomly resampled ovary data and cyclic data. On the contrary, we estimated the number of clusters to be around four to six and around six to seven on the mixture of normal and randomly resampled ovary data respectively using the FOM approach. Therefore, although the FOM approach cannot give definite answers to the correct number of clusters, it still gives a reasonable hint to the number of clusters relative to the silhouette approach.

Table 3.2: The number of clusters that maximizes the average silhouette width for each of the six clustering algorithms on real and synthetic data. The number of clusters that achieves the highest average silhouette width over all clustering algorithms for each data set is shown in bold.

data	*clustering algorithms*					
(# classes)	CAST	k-means	single-link	average-link	complete-link	centroid-link
ovary (4)	**6**	7	2	7	6	3
cell cycle (5)	**5**	5	2	5	5	6
rat (4)	3	5	2	**2**	30	**2**
normal (4)	2	**2**	2	2	2	2
resampled (4)	**2**	**2**	2	2	**2**	2
cyclic (10)	3	**2**	2	3	2	3

3.5 Conclusions and Future Work

3.5.1 Summary and Conclusions

The main contribution of this chapter is not the comparison of specific algorithms on specific data sets, but rather the development of a simple, quantitative data-driven methodology

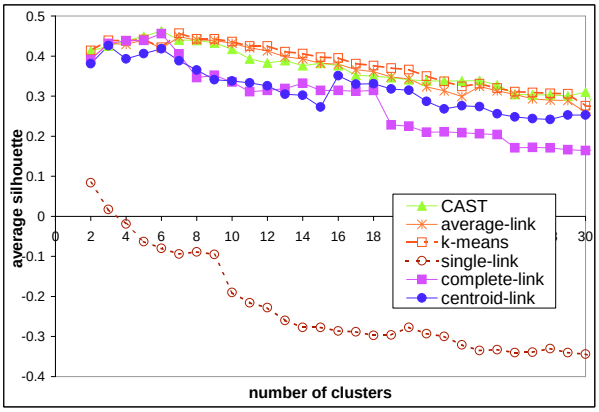

Figure 3.13: Average silhouette widths over different numbers of clusters on the ovary data.

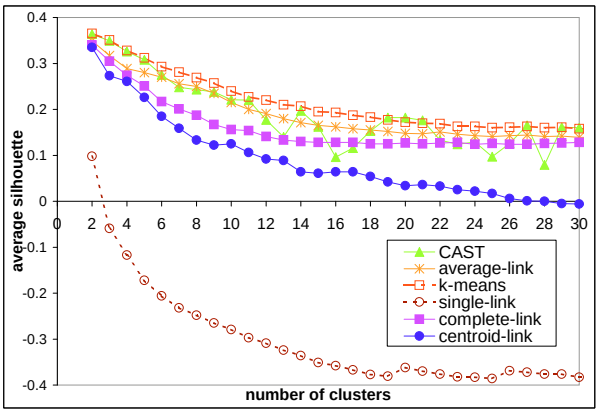

Figure 3.14: Average silhouette widths over different numbers of clusters on the mixture of normal data.

allowing such comparisons to be made between any clustering algorithms on any data set. We presented experimental evidence that our FOM methodology produces results that are well correlated with biologically relevant external standards on real data sets and with the artificial external criteria on synthetic data sets (external validation from Section 3.4.1).

Although comparison between specific clustering algorithms is not our primary focus, we presented comparisons between several important clustering algorithms over different data sets. Our results in Section 3.3 confirmed the general belief that average-link and complete-link algorithms tend to produce higher quality clusters than single-link. We also showed that CAST tends to have relatively high predictive power. Furthermore, we showed that the iterative k-means step after average-link improves cluster quality since the FOM's

after the additional iterative k-means step tend to be lower. Based on our results, we would recommend using CAST or k-means for analysis of gene expression data, and would recommend against using single-link.

In addition, we compared the results from our FOM approach to other internal validation approaches that do not require any external knowledge of the data in Section 3.4.2. Specifically, we compared the FOM approach to the silhouette approach by Rousseeuw [68] and the homogeneity and separation approach by Shamir and Sharan [75]. Our results from Section 3.4.2 showed that the FOM approach tends to have higher agreement to external validation results than the other two approaches. Furthermore, although the adjustment factor, which was derived in Section 3.1.3 to compensate for the statistical bias of increasing the number of clusters, is not perfect in the FOM approach and the FOM approach can only approximate the correct number of clusters within a range, we showed that the silhoutte approach does not excel over our FOM approach in prediction of the correct numbers of clusters.

3.5.2 Limitations and Future Work

Our FOM methodology takes a *predictive* approach, *i.e.*, our model assumes that the left-out experiment contains information from the experiments that are used to produce clusters. In other words, our approach compares the relative strength in predictive power of clustering algorithms given the related information in the experiments used to produce clusters. Our approach is not applicable to all situations: if all experiments contain independent information, no predictive approach is possible. Despite the limitations, we believe that our FOM method is applicable to many gene expression data sets. We successfully applied our method to data sets with varying degree of dependence including time series data (the yeast cell cycle data [20] and the rat data [81]) and data sets with different types of tissue samples (the ovary data [71]).

Another limitation of the FOM approach is that it is not easy to compare clustering results with different numbers of clusters. This is because the adjustment factor we derived in Section 3.1.3 is not perfect, and the adjusted FOM's decrease as the number of clusters

is increased on all of our data sets.

In this chapter, we used the same similarity metric in all of our experiments. We are not sure if our definition of FOM is biased toward any particular similarity metric. Therefore, we recommend comparing all clustering results generated using the same similarity metric. An interesting direction of further research would be to investigate the effect of different similarity metrics on our definition of the FOM, and possible alternate definitions of FOM that depend on the similarity metrics used in clustering algorithms. For example, if the goal is to capture anti-correlated genes and the absolute value of the correlation coefficient is used to compute pairwise similarities between genes, anti-correlated gene clusters would yield high FOM's with our current definition of FOM that measures within-cluster variation. Hence, the current definition of FOM is not appropriate in this case.

The nature of our methodology to leave out each experiment in turn and repeat clustering makes it computationally intensive for large data sets with lots of experiments. A direction of future work is to leave out groups of experiments at a time for large data sets.

Another direction of future work is to compare our FOM predictive approach to other measures of cluster validation in addition to the silhouette approach and the homogeneity and separation measures. Toldo [80] recently adopted a linear algebra approach from Mather [57] to validate clustering results from gene expression data. It would be interesting to compare our FOM approach with this linear algebra approach.

To summarize, clustering is a difficult problem. We believe that the methodology introduced in this chapter for quantitative comparison of the predictive power of clustering algorithms will prove to be a valuable ingredient in future clustering studies.

Chapter 4

PRINCIPAL COMPONENT ANALYSIS FOR CLUSTERING GENE EXPRESSION DATA

In the clustering literature, dimension reduction techniques are sometimes applied to the data before cluster analysis. The hope is that the transformed and reduced set of variables captures the essence of the data while reducing the noise level. Gene expression data are expected to be noisy, and there are usually experiments containing duplicate information, like the same type of tissue samples or consecutive time points. For example, the ovary data described in Chapter 2 consists of many experiments using normal ovarian tissue samples. In this chapter, we empirically investigate the effectiveness of a dimension reduction technique called *principal component analysis* (PCA) in clustering gene expression data.

4.1 Principal Component Analysis (PCA)

4.1.1 An Example of PCA

The central idea of principal component analysis (PCA) is to reduce the dimensionality of the data set while retaining as much as possible the variation in the data set. Principal components (PC's) are linear transformations of the original set of variables. PC's are uncorrelated and ordered so that the first few PC's contain most of the variations in the original data set [43].

The first PC has the geometric interpretation that it is a new coordinate axis that maximizes the variation of the projections of the data points on the new coordinate axis. Figure 4.1 shows a scatterplot of some fictitious data points in two dimensions (x_1 and x_2). The points show an elliptical shape, and the first PC is in the direction of the principal axis of this ellipse (marked PC_1 in Figure 4.1). The second PC is orthogonal to the first PC and is marked PC_2 in Figure 4.1. If the data points are projected onto the first PC, most

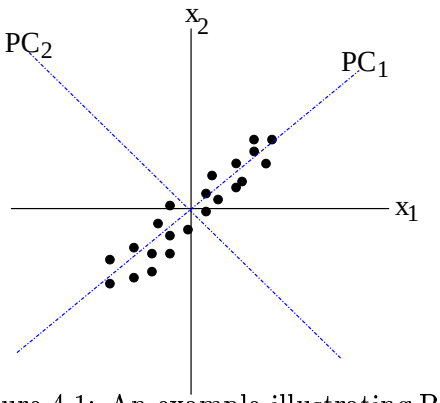

Figure 4.1: An example illustrating PCA

of the variation of the two dimensional data points would be captured in one dimension.

4.1.2 Definitions of PCA

Let X be a gene expression data set with n genes and p experiments. Our goal is to cluster the genes in the data set, so the experiments are the variables. Let $\bar{\mathbf{x}}_\mathbf{j}$ be a column vector of the expression levels of all n genes under experiment j. A PC is a linear transformation of the original variables (experiments). Let $\bar{\mathbf{z}}_\mathbf{k} = \sum_{j=1}^{p} \alpha_{k,j}\bar{\mathbf{x}}_\mathbf{j}$ be the kth PC, where the $\alpha_{k,j}$'s are scalars. In particular, the first PC, $\bar{\mathbf{z}}_\mathbf{1}$, can be written as $\sum_{j=1}^{p} \alpha_{1,j}\bar{\mathbf{x}}_\mathbf{j}$. Let Σ be the covariance matrix of the data, *i.e.*, $\Sigma(i,j)$ is the covariance between experiment i and experiment j where $i \neq j$ and $\Sigma(i,i)$ is the variance of experiment i. Suppose $\bar{\alpha}_\mathbf{k}$ is a column vector of all the $\alpha_{k,j}$'s, *i.e.*, $\bar{\alpha}_\mathbf{k}^\mathbf{T} = (\alpha_{k,1}, \alpha_{k,2}, \ldots, \alpha_{k,p})$. The first PC captures the maximum amount of variation in the data. To derive the first PC, we have to find $\bar{\alpha}_\mathbf{1}$ that maximizes $var(\bar{\mathbf{z}}_\mathbf{1}) = var(\sum_{j=1}^{p} \alpha_{1,j}\bar{\mathbf{x}}_\mathbf{j}) = \bar{\alpha}_\mathbf{1}^\mathbf{T}\Sigma\bar{\alpha}_\mathbf{1}$, subject to the constraint $\bar{\alpha}_\mathbf{1}^\mathbf{T}\bar{\alpha}_\mathbf{1} = 1$. It can be shown that $\bar{\alpha}_\mathbf{1}$ is the eigenvector corresponding to the largest eigenvalue, λ_1, of Σ, and $var(\bar{\mathbf{z}}_\mathbf{1}) = \lambda_1$ [43]. In general, the kth PC, $\bar{\mathbf{z}}_\mathbf{k} = \sum_{j=1}^{p} \alpha_{k,j}\bar{\mathbf{x}}_\mathbf{j}$, can be derived by maximizing $var(\sum_{j=1}^{p} \alpha_{k,j}\bar{\mathbf{x}}_\mathbf{j})$, such that $\bar{\alpha}_\mathbf{k}^\mathbf{T}\bar{\alpha}_\mathbf{k} = 1$ and $\bar{\alpha}_\mathbf{k}^\mathbf{T}\bar{\alpha}_\mathbf{i} = 0$, where $i < k$. It can be shown that $\bar{\alpha}_\mathbf{k}$ is an eigenvector of Σ corresponding to its kth largest eigenvalue λ_k, and $var(\bar{\mathbf{z}}_\mathbf{k}) = \lambda_k$ [43].

In the case of gene expression data, the population covariance matrix Σ is not known. The sample covariance matrix S can be used instead. Let $x_{i,j}$ be the gene expression level

of gene i under experiment j. The sample covariance between experiments j and k, $S(j,k)$, can be calculated as $\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j}-\mu_{x_j})(x_{i,k}-\mu_{x_k})$, where $\mu_{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{i,j}$.

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 in applying SVD to gene expression data, for example, Holter *et al.* [38] and Alter *et al.* [3].

From the derivation of 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, and the kth PC has the kth largest variance among all PC's. Since most of the variation of high dimensional data points can be captured in reduced dimensions defined by the first few PC's, the first few PC's are often used in visualization of high dimensional data points.

4.1.3 Choosing the number of PC's

Since the variance of the PC's are ordered, usually the first m ($m \leq p$, where p is the number of experiments in the data) PC's are used in data analysis. The next question is how we should choose m, the number of first PC's to be retained, to adequately represent the data set. There are some common rules of thumb to choose the number of components to retain in PCA. Most of the rules are informal and *ad hoc*. The first common rule of thumb is to choose m to be the smallest integer such that a chosen percentage of total variation is exceeded. Another common approach uses a *scree graph*, in which the kth eigenvalue is plotted against the component number, k. The number of components m is chosen to be the point at which the line in the scree graph is "steep" to the left but "not steep" to the right. The main problem with these approaches is that they are very subjective. There are some more formal approaches in the literature, but in practice, they tend not to work as well as the *ad hoc* approach [43].

4.1.4 Covariance versus correlation matrices

In the PCA literature, some authors prefer to define PC's using the *correlation* matrix instead of the covariance matrix. The correlation between a pair of variables is equivalent to the covariance divided by the product of the standard deviations of the two variables. Extracting the PC's as the eigenvectors of the correlation matrix is equivalent to computing the PC's from the original variables after each has been standardized to have unit variance. PCA based on covariance matrices has the potential drawback that the PC's are highly sensitive to the unit of measurement. If there are large differences between the variances of the variables, then the first few PC's computed with the covariance matrix are dominated by the variables with large variances. On the other hand, defining PC's with the correlation matrix has the drawback that the data is arbitrarily re-scaled to have unit variance. The general rule of thumb is to define PC's using the correlation matrix if the variables are of different types [43]. We assume the raw expression levels from different experiments (variables) have been scaled (for example, each experiment has been scaled to have the same average intensity), and hence we compute PC's from the covariance matrix.

4.1.5 Application of PCA in cluster analysis

In the clustering literature, PCA is sometimes applied to reduce the dimension 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 variation in the data, are usually used in cluster analysis (for example, Jolliffe *et al.* [44]). Figure 4.2a is a fictitious situation in which applying PCA before cluster analysis may help. The first PC (dotted line) is in the direction of inter-cluster separation in Figure 4.2a. Projection of the data points on the first PC clearly highlights the separation between the two clusters in the data. However, PCA does not help in all situations. For example, in Figure 4.2b, the first PC is in the direction of x_2. Projection of the data points onto the first PC destroys the separation between the two clusters in the data.

In addition to the fictitious examples illustrating the possible pros and cons of PCA on

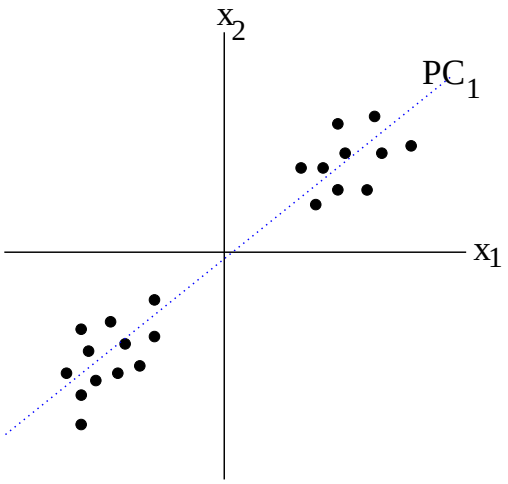

Figure 4.2a: An example in which the first PC captures the cluster structure.

cluster analysis, Chang [17] showed theoretically that the first few PC's may not contain cluster information under certain assumptions. Assuming that the data is a mixture of two multivariate normal distributions with different means but with an identical within-cluster covariance matrix, Chang [17] derived a relationship between the distance of the two subpopulations and any subset of PC's, showing that the set of PC's with the largest eigenvalues does not necessarily contain more cluster structure information (the distance between the two subpopulations is used as a measure of discriminatory power for cluster structures). He also generated an artificial example in which there are two classes, and if the data points are visualized in two dimensions, the two classes are only well-separated in the subspace of the first and last PC's.

There are two popular similarity metrics in clustering gene expression data: Euclidean distance and correlation coefficient (see Chapter 2). The pairwise Euclidean distance between two objects is unchanged after the PCA step if all p PC's are used. When Euclidean distance is used as the similarity metric, using the first m PC's simply provides an approximation to the similarity metric [43]. When correlation coefficient is used as the similarity metric, the pairwise correlation coefficient between two objects is not the same after the PCA step even if all p PC's are used. There is no simple relationship between the correlation

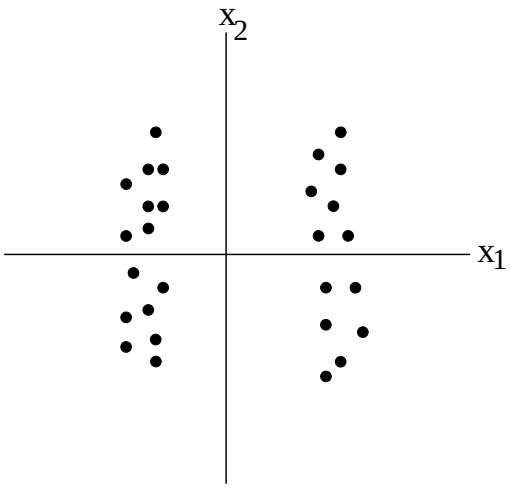

Figure 4.2b: An example in which the first PC does not capture the cluster structure.

coefficients of the same pair of objects with and without PCA. In general, the extra computation to find the PC's far outweighs any reduction in running time for using fewer PC's to compute the pairwise similarity. So, the hope for using PCA prior to cluster analysis is to improve the quality of clustering results, and not to reduce computational time.

4.2 Motivation

PCA has been applied in the context of gene expression analysis to visualize and identify clusters. For example, in Raychaudhuri *et al.* [67], PCA was applied to the sporulation data set [21].[1] The sporulation data set shows the temporal expression patterns of approximately 6000 yeast genes over seven successive time points. Chu *et al.* [21] identified seven clusters in a subset of the sporulation data set (477 genes). Figure 4.3a 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 and show a unimodal distribution in Figure 4.3a. From the visualization of the data in the space of the first two PC's, Raychaudhuri *et al.* [67]

[1]Sporulation is the process in which diploid cells undergo meiosis to produce haploid cells in reproduction of yeast.

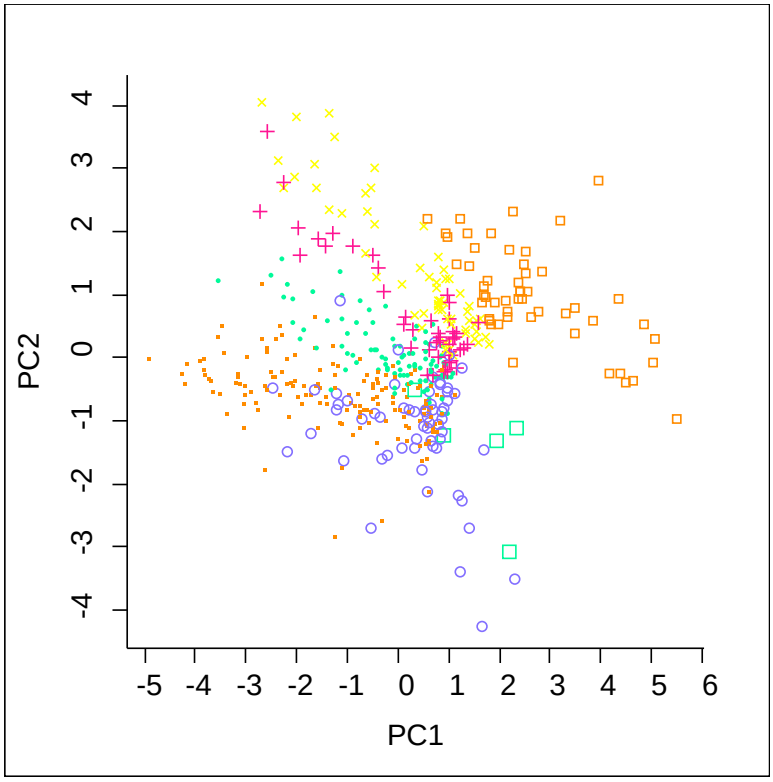

Figure 4.3a: Visualization of a subset of the sporulation data in the space of the first two PC's.

concluded that the data may not contain any clusters. However, if we view the same subset of data points in the space of the first 3 PC's (containing 93.2% of the variation in the data) in Figure 4.3b, the seven patterns are much more separated. This example shows that a small variation (7.3%) 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.

With Chang's theoretic results and the possibility of the situation in Figure 4.3 in mind, it is clear that clustering with the PC's instead of the original variables does not have universal success. However, the theoretical results in Chang [17] are true only under an unrealistic assumption for gene expression data (*i.e.,* there are two classes and each of the classes is generated according to the multivariate normal distribution with a common covariance matrix). Therefore, there is a need to investigate the effectiveness of PCA as a

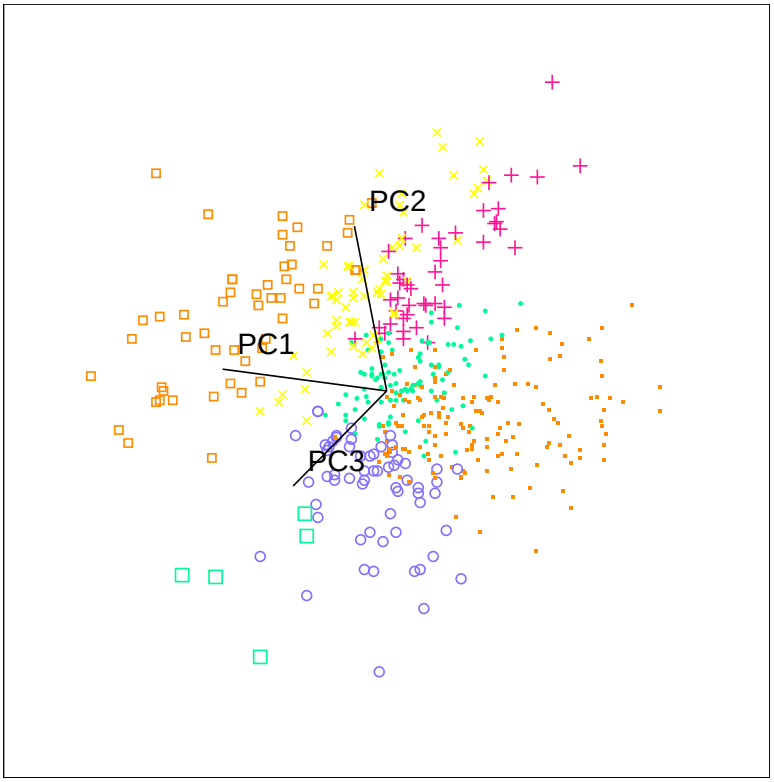

Figure 4.3b: Visualization of a subset of the sporulation data in the space of the first three PC's.

preprocessing step to cluster analysis on gene expression data before any conclusions are drawn. Specifically, we would like to address the following questions:

1. Is the traditional wisdom of using the first few PC's in cluster analysis a good strategy? If so, how many first PC's should we use?

2. Does there exist a set of PC's that produces the best possible cluster quality? If so, how does the best possible cluster quality compare with the quality of clustering results from using the original data and from using the first few PC's?

3. Is there a pattern for the PC's that produce the best quality clusters? If not, how does the quality of clusters produced by random subsets of PC's compare with the quality of clusters produced by the first PC's and the best quality clusters?

4. Is there anything special about clustering with subsets of PC's?

   (a) PC's are special cases of orthogonal bases that are ordered with respect to variations in the original data. How does the quality of clustering results from using subsets of orthogonal vectors (that are not ordered) compare with the quality of clustering results from using subsets of PC's?

   (b) How about subsets of columns of the original data? In particular, are there subsets of columns of the original data that would produce high quality clustering results?

4.3 Overview of Our Methodology

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 subsets of PC's or orthogonal vectors. In order to address the questions raised in Section 4.2, we need a method to evaluate the quality of clustering results. The effectiveness of clustering with the original data and with different subsets of PC's or orthogonal vectors is determined by comparing the clustering results to an objective external criterion of the data using the adjusted Rand index described in Chapter 2. Although external criteria are rarely available for gene expression data in general, we hope to gain some insights into the quality of results from clustering with different subsets of PC's or orthogonal vectors by using both real and synthetic data sets for which external criteria are known.

4.3.1 Subsets of PC's

Motivated by Chang's theoretical result [17], 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

The greedy approach

- Exhaustive search for a set of m_0 PC's with maximum adjusted Rand index. Denote this 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}.

Figure 4.4: Pseudo-code for the greedy approach.

index is by exhaustive search over all possible sets of PC's. However, exhaustive search is computationally very expensive. 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 [1]. Let m_0 be the minimum number of PC's to be clustered, and p be the number of experiments in the data. This approach starts with an exhaustive search for a set of m_0 PC's with maximum adjusted Rand index. PC's are added greedily to the current set one at a time, such that the expanded set (the additional PC together with the current set) gives the maximum adjusted Rand index when the data with all the genes under this expanded set of components is clustered. The pseudo-code of the greedy algorithm is given in Figure 4.4.

The modified greedy approach: The modified greedy approach requires an additional integer parameter, r, which represents the number of *best* solutions to keep in each search step. Denote the best r sets of components as $\mathcal{S}_m = \{S_m^1, \ldots, S_m^r\}$, 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, r sets of components that achieve the top r adjusted Rand indices are stored. For each m (where $m = (m_0+1), \ldots, p$) and each of the S_m^i (where $i = 1, \ldots, r$), 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 r 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 $r = 1$, the modified greedy approach is identical to the simple greedy approach, and when $r = \binom{p}{m}$, the modified greedy approach is reduced to exhaustive search. So the choice for r is a tradeoff between running time and quality of solution. In our experiments, r was set to be 3.

Random PC's: We also investigated the effect on the quality of clusters obtained from random subsets of PC's (in contrast to subsets of PC's obtained from the greedy and modified greedy approaches). Multiple random sets of PC's (30 in our experiments) were chosen to compute the average and standard deviation of the adjusted Rand indices.

4.3.2 Orthogonal bases

We would like to evaluate the quality of clustering results when data points are projected onto the subspaces defined by orthogonal bases. An orthogonal basis is a set of linear independent and orthogonal vectors that spans the subspace of the p original column vectors. Unlike the PC's, there is no unique orthogonal basis spanning the subspace and the vectors from an orthogonal basis are not ordered. We generated many different random orthogonal bases, and applied both the greedy approach and the modified greedy approach from Section 4.3.1 to search for orthogonal bases with "high" adjusted Rand indices. Furthermore, we computed the adjusted Rand indices of random subsets of vectors from the orthogonal bases to compare with random subsets of PC's.

4.3.3 Original data

As a control, we also evaluated the quality of clustering results from different subsets of columns of the original data. We applied both the greedy and modified greedy approaches from Section 4.3.1 to search for subsets of columns of the original data with "high" adjusted Rand indices.

4.3.4 Summary

Given a gene expression data set with n genes and p experiments with an external criterion, our evaluation methodology consists of the following steps:

- Apply clustering algorithm A to the given data set, and compute the adjusted Rand index with the external criterion.
- Apply PCA to the data
    1. Apply clustering algorithm A to the first m PC's (where $m = m_0, \ldots, p$), and compute the adjusted Rand index for each of the clustering results.
    2. Select subsets of m PC's (where $m = m_0, \ldots, p$) with high adjusted Rand indices using the greedy and modified greedy approaches on the clustering results from algorithm A.
    3. Cluster random subsets of PC's with algorithm A.
- Generate random orthogonal bases from the given data set. For each random orthogonal basis:
    1. Apply the greedy and the modified greedy approaches to determine subsets of m orthogonal vectors (where $m = m_0, \ldots, p$) with high adjusted Rand indices from algorithm A.
    2. Cluster random subsets of the orthogonal basis with algorithm A.
- Apply the greedy and modified greedy approaches to the original data to search for subsets of m variables (where $m = m_0, \ldots, p$) with "high" adjusted Rand indices.

4.4 Experimental details

4.4.1 Data sets

Both real gene expression data sets with external criteria and synthetic data sets are used in this empirical study. Specifically, we used the ovary data and the yeast cell cycle data with the 5-phase criterion described in Section 2.3.1. In addition, we illustrated our approach with the subset (477 genes) of sporulation data that motivated our empirical study (see Section 4.2). We also complemented our empirical study with all three sets of synthetic data described in Section 2.3.2.

4.4.2 PCA on the data sets

We applied PCA to both real and synthetic data sets. The scree graphs for the ovary data and yeast cell cycle data are shown to illustrate the amounts of variation captured by different PC's. In a scree graph, eigenvalues, which are proportional to the variance of the PC's, are plotted against the number of PC's. From the scree graph for the ovary data in Figure 4.5, there is a sharp drop of steepness (*i.e.*, amount of variation) at the third PC, and another gentle change of steepness at the sixth PC. The first 14 PC's account for over 90% of the total variation of the ovary data. Similarly, there is a sharp drop in the amount of variation at the third PC, and another gentle drop at the fifth PC in the scree graph for the yeast cell cycle data shown in Figure 4.6. The first 8 PC's account for over 90% of the total variation of this data.

4.4.3 Clustering algorithms and similarity metrics

In our experiments, we assume the number of clusters is known and clustering results with the correct number of clusters are produced. We used three clustering algorithms in our empirical study: the Cluster Affinity Search Technique (CAST), the hierarchical average-link algorithm, and the k-means algorithm. These algorithms are described in detail in Section 2.1.3.

In our experiments, we evaluated the effectiveness of PCA in cluster analysis with both Euclidean distance and correlation coefficient, namely, CAST with correlation coefficient,

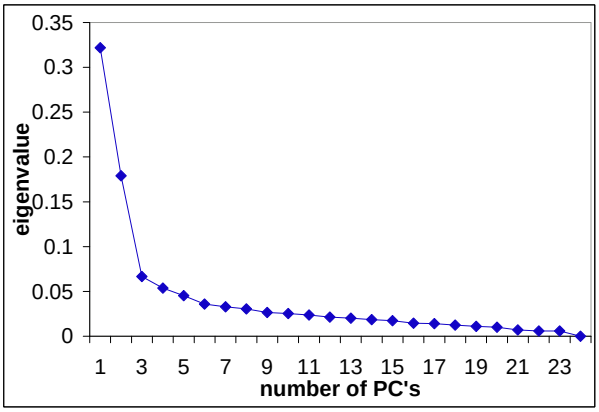

Figure 4.5: Scree graph for the ovary data

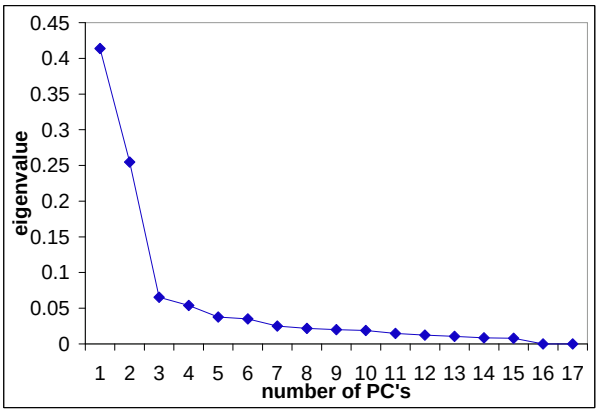

Figure 4.6: Scree graph for the yeast cell cycle data

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, see Appendix B for the proof).

4.5 Results

Here are the overall conclusions from our empirical study:

1. The quality of clustering results (*i.e.*, the adjusted Rand index with the external criterion) using the first PC's is not necessarily higher than that on the original data. There is no obvious relationship between cluster quality and the number of first PC's used.

2. In most cases, there exists another set of m PC's (determined by the greedy or modified greedy approaches) that achieves a higher adjusted Rand index than the first m components (where $m = m_0, \ldots, p$).

3. Random subsets of PC's usually produce much lower average adjusted Rand indices than clustering with the first PC's, especially when the number of PC's is small.

4. There exist subsets of vectors from the original data set and from random orthogonal bases that achieve comparable adjusted Rand indices to subsets of PC's (determined by the greedy and modified greedy approaches) when the number of components is large. However, when the number of components is small, the adjusted Rand indices from the greedy and modified greedy approaches on random orthogonal bases tend to be lower than those from subsets of PC's.

5. On average, the quality of clusters obtained by clustering random subsets of PC's tend to be slightly lower than those obtained by clustering random subsets of orthogonal bases, 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 data (without PCA), while the adjusted Rand indices computed using PC's vary with the number of components. The results using the hierarchical average-link clustering algorithm show similar patterns to the results using k-means (but with slightly lower adjusted Rand indices), and hence are not shown.

4.5.1 Gene expression data

The ovary data

CAST: Figure 4.7a 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 the adjusted Rand index using the original data (no 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 data. Moreover, the adjusted Rand indices computed using the greedy and modified greedy approaches are not very different. Figure 4.7b shows the additional results of the average adjusted Rand indices of random sets of PC's and random projections from orthogonal bases. The standard deviation in the adjusted Rand indices of the multiple runs (30) of random subsets of orthogonal projections are represented by the error bars in Figure 4.7b. 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 4.7b 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.

Figure 4.7c shows the results of the greedy and modified greedy approaches on the original data and two different random orthogonal bases. When the number of components is small (below 7), the adjusted Rand indices from the greedy and modified greedy approaches on both the original data and the random orthogonal bases are lower than those from the PC's. This shows that there exists a few good PC's (not necessarily the first PC's) giving high adjusted Rand indices. However, there is no well-established method to determine a

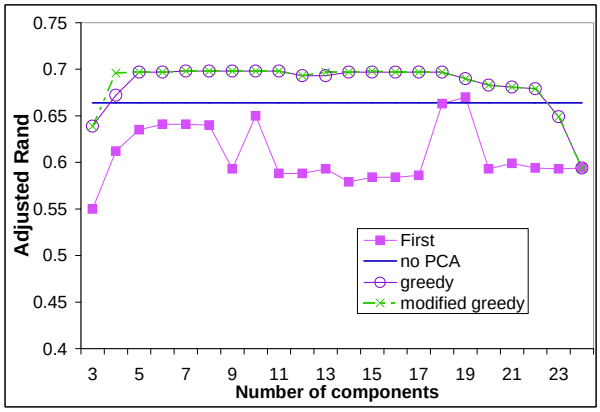

Figure 4.7a: Adjusted Rand index against the number of components using CAST and correlation on the ovary data.

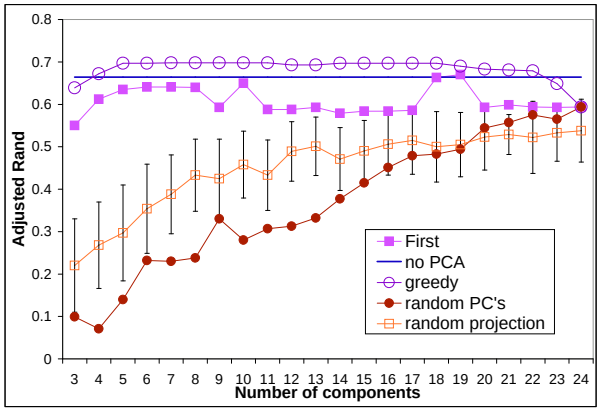

Figure 4.7b: Adjusted Rand index against the number of components using CAST and correlation on the ovary data (with random PC's and random orthogonal bases).

set of PC's capturing the cluster structure without using the external criterion. When the number of components is large, the adjusted Rand indices from the greedy and modified greedy approaches on the original data and the random orthogonal bases are comparable to those from the PC's. The trend shown in Figure 4.7c is typical for applying greedy and modified greedy approaches to the original data and random orthogonal bases of other data sets, so only the results of the greedy and modified greedy approaches on the PC's are shown in the forthcoming sections.

K-means: Figure 4.7d and Figure 4.7e show the adjusted Rand indices using the k-means algorithm on the ovary data with correlation and Euclidean distance as similarity metrics

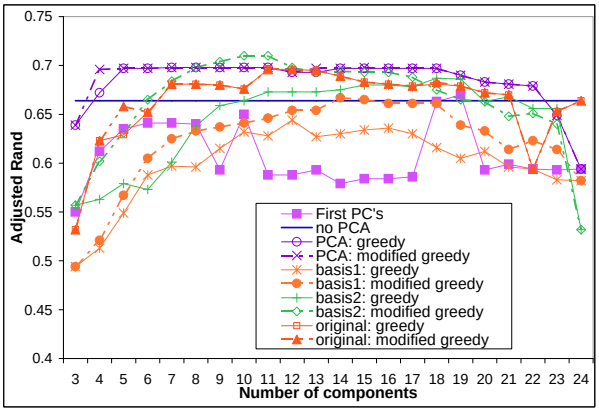

Figure 4.7c: Adjusted Rand index against the number of components using CAST and correlation on the ovary data (with results applying the greedy and modified greedy approaches to the original data and random orthogonal bases).

respectively. Figure 4.7d shows that the adjusted Rand indices using the first m components tend 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 4.7e): 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. Figure 4.7d and Figure 4.7e 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 4.7d show an irregular

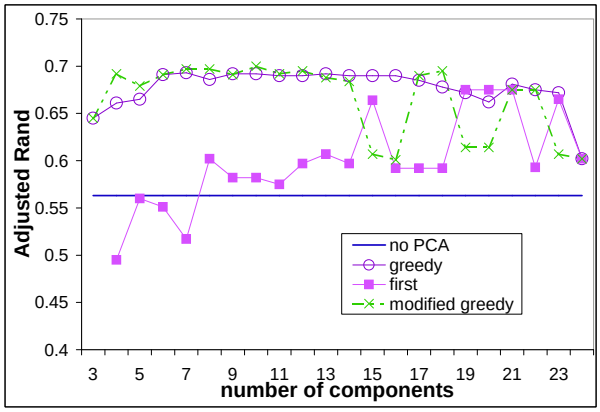

Figure 4.7d: Adjusted Rand index against the number of components using k-means and correlation on the ovary data.

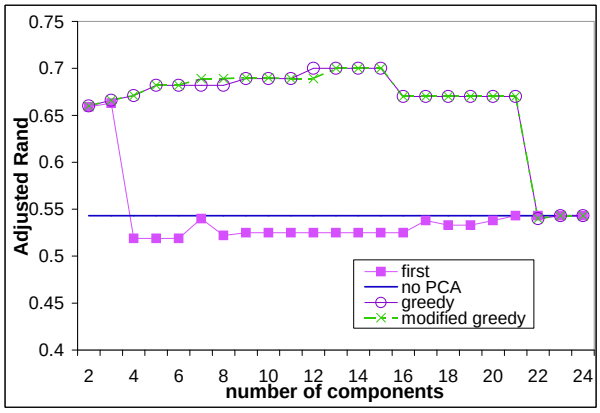

Figure 4.7e: Adjusted Rand index against the number of components using k-means and distance on the ovary data.

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 4.7a and 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 4.7d), 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.

The yeast cell cycle data

CAST: Figure 4.8a 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 4.8b 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 the first few PC's. For this data, the first 8 PC's cover over 90% of the total variation in the data. Clustering this data with the first 8 PC's produces lower quality results than clustering with the original data. Figure 4.8b also shows a very different picture than Figure 4.7e on the ovary data: there is no clear pattern for using different numbers

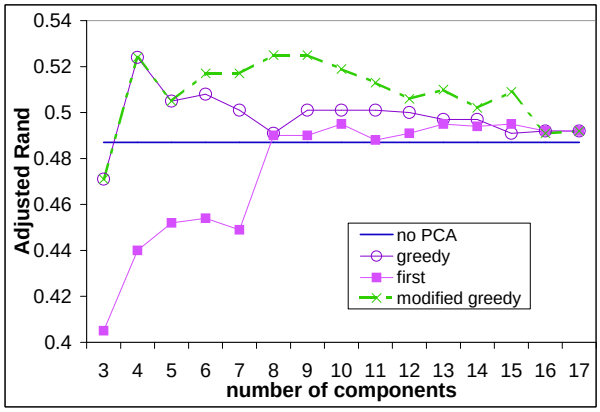

Figure 4.8a: Adjusted Rand index against the number of components using CAST and correlation on the yeast cell cycle data.

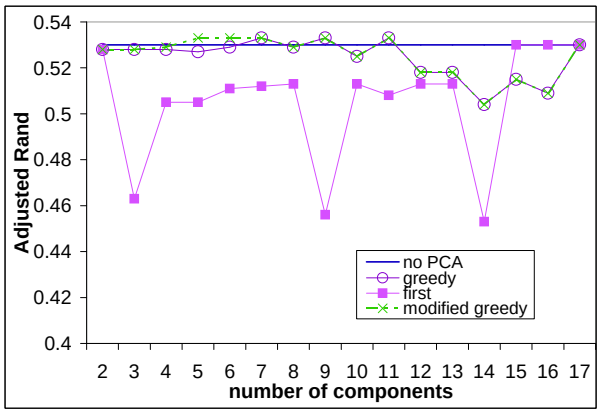

Figure 4.8b: Adjusted Rand index against the number of components using k-means and distance on the cell cycle data.

of first PC's, and the greedy and modified greedy approaches did not find subsets of PC's that result in higher quality clusters. This shows that the effectiveness of PC's in capturing the cluster structure depends on the data set.

The results on the yeast cell cycle data using k-means with correlation are similar to those using k-means and Euclidean distance (figure not shown), except the adjusted Rand indices of the modified greedy approach are much higher than those of the greedy approach.

The sporulation data

CAST: Figure 4.9a shows the results using CAST and correlation on the subset of sporulation data with 477 genes (the same subset as in Figure 4.3). The results show that clustering with the original data gives much higher adjusted Rand indices than clustering with the first PC's no matter how many PC's are used. Figure 4.9a also shows a surprising result: the greedy and the modified greedy approaches cannot find a set of PC's that would excel over clustering with the original data.

K-means: The results using k-means and Euclidean distance on the subset of sporulation data in Figure 4.9b show a very different picture than when CAST and correlation is used: clustering with the first PC's achieves comparable adjusted Rand indices to clustering with the original data. However, the adjusted Rand indices (approximately 0.23) when k-means and Euclidean distance are used are much lower than clustering the original data with CAST and correlation (0.39). The k-means algorithm with correlation did not converge on the sporulation data, so the results are not shown.

These results show that CAST and correlation produces much higher cluster quality than k-means and Euclidean distance on the sporulation data. Moreover, clustering with the first PC's produces clustering results with relatively low adjusted Rand indices (below 0.25 for all three clustering algorithms with either Euclidean distance or correlation). Hence, the first PC's probably do not capture the cluster structure in this data.

4.5.2 Synthetic data

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 4.5.1. Figure 4.10a 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 4.10a represent the average adjusted Rand indices over the ten 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

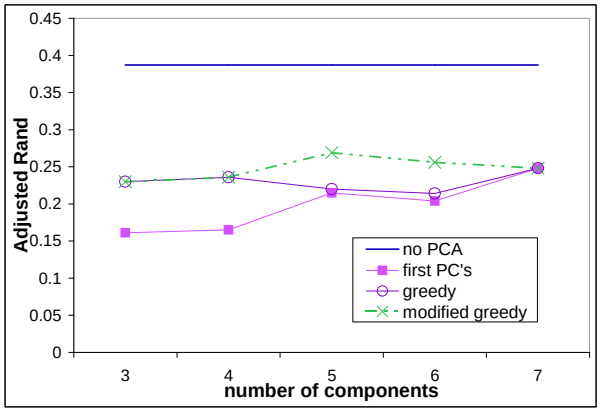

Figure 4.9a: Adjusted Rand index against the number of components using CAST and correlation on the sporulation data.

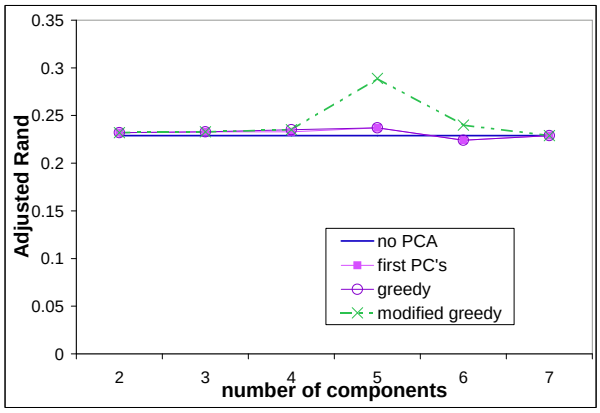

Figure 4.9b: Adjusted Rand index against the number of components using k-means and distance on the sporulation data.

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. Based on the Wilcoxon signed rank test [36], the adjusted Rand index without PCA is greater than that with the first m components at the 5% significance level for $m = 3, \ldots, 21$. A manual study of the experimental results from each of the ten replicates (details not shown here) shows that eight out of the ten

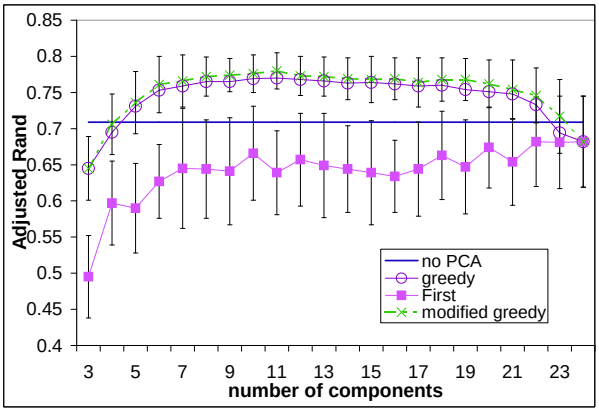

Figure 4.10a: Average adjusted Rand index against the number of components using CAST and correlation on the mixture of normal synthetic data.

replicates show very similar patterns to the average pattern in Figure 4.10a, *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 4.10a shows a similar trend to the ovary data in Figure 4.7a, 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 4.10b and Figure 4.10c respectively. In Figure 4.10b, the average adjusted Rand indices using the first m components gradually increase as the number of components increases. Based on the Wilcoxon signed rank test, 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 4.10c, 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.

Randomly resampled ovary data

Figure 4.11a and Figure 4.11b show the average adjusted Rand indices using CAST with correlation, and k-means with Euclidean distance on the randomly resampled ovary data.

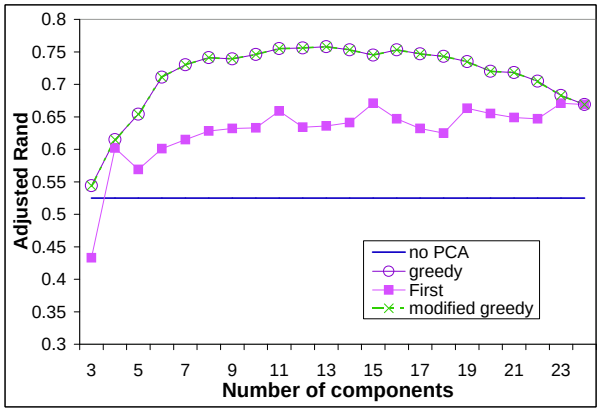

Figure 4.10b: Average adjusted Rand index against the number of components using k-means and correlation on the mixture of normal synthetic data.

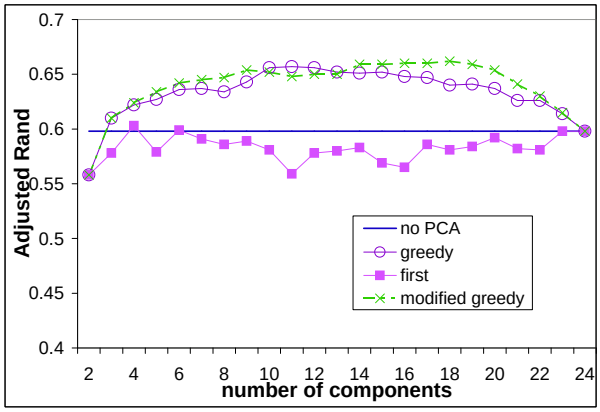

Figure 4.10c: Average adjusted Rand index against the number of components using k-means and distance on the mixture of normal synthetic data.

The general trend is very similar to the results on the ovary data and the mixture of normal distributions.

Cyclic data

Figure 4.12a 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 4.12b 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

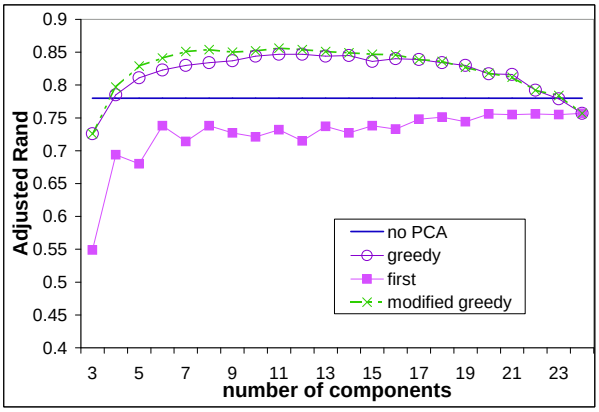

Figure 4.11a: Average adjusted Rand index against the number of components using CAST and correlation on the randomly resampled synthetic data.

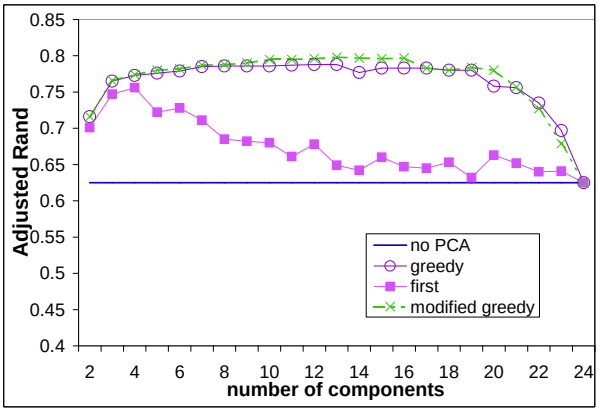

Figure 4.11b: Average adjusted Rand index against the number of components using k-means and distance on the randomly resampled synthetic data.

with the first PC's is not higher than that from clustering with the original variables.

4.5.3 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 are usually higher than the adjusted Rand index from clustering with the original variables.

Table 4.1 summarizes the results of comparing the adjusted Rand indices from clustering with the first PC's to the adjusted Rand indices from clustering the original real gene expression data. Each entry in Table 4.1 shows the fraction of times for which the adjusted

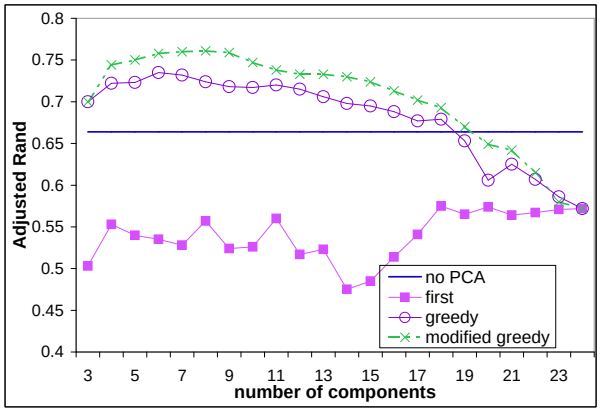

Figure 4.12a: Average adjusted Rand index against the number of components using CAST and correlation on the cyclic synthetic data.

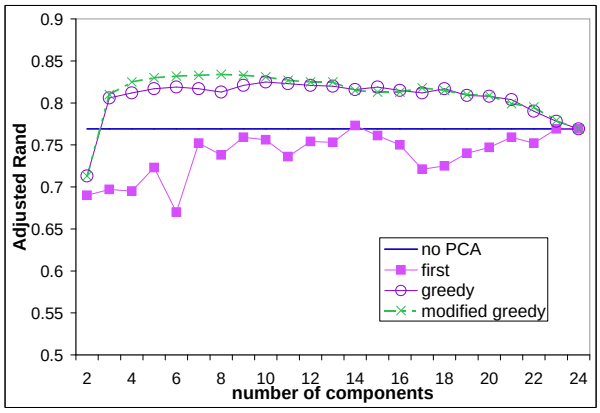

Figure 4.12b: Average adjusted Rand index against the number of components using k-means and distance on the cyclic synthetic data.

Rand indices from clustering with the first PC's are *greater* than that from the original data. Specifically, the denominator in each entry shows the total number of components considered, *i.e.*, $p - m_0 + 1$ minus the number of unavailable clustering results. A clustering result may be unavailable because the algorithm does not converge. The numerator shows the number of times that the adjusted Rand indices from clustering with the first PC's are greater than that from clustering the original data. Hence, a small fraction in Table 4.1 means that clustering with the first PC's produces relatively low adjusted Rand indices in most cases. For example, on the ovary data, the total number of components considered (denominator) is $24 - 3 + 1 = 22$ when correlation is used as the similarity metric, and is

Table 4.1: Summary of comparing the adjusted Rand indices from clustering with the first components to those from clustering with the original real gene expression data. Each entry shows the fraction of times for which the adjusted Rand indices from clustering with the first PC's are *greater* than that from the original data.

data	*CAST* *correlation*	*k-means* *correlation*	*k-means* *distance*	*average-link* *correlation*	*average-link* *distance*
ovary data	1/22	17/21	2/23	18/22	2/23
cell cycle data	10/15	1/12	0/16	2/15	4/16
sporulation	0/5	NA	4/6	0/5	3/6

$24-2+1=23$ when Euclidean distance is used assuming all clustering results are available. The entry 17/21 under the k-means algorithm using correlation as the similarity metric on the ovary data means that out of the 21 available clustering results, clustering with the first PC's produces higher adjusted Rand indices than clustering with the original data 17 times. In general, Table 4.1 shows that clustering with the first PC's often produces *lower* adjusted Rand indices than clustering with the original real data. On the sporulation data, clustering with the first PC's tends to produce higher adjusted Rand index than clustering with the original data when k-means and average-link are applied with Euclidean distance. However, in both cases, the differences in adjusted Rand indices are very small (the average absolute differences in adjusted Rand indices are 0.004 and 0.02 for k-means and average-link respectively).

A rule of thumb for choosing the number of first PC's is the minimum number of first PC's that accounts for at least 90% of the total variations in the data. Table 4.2 shows the results of comparing the adjusted Rand index from clustering with the number of first PC's that covers 90% of the variation on each real data set to that from clustering with the original data. An entry marked "+" means that clustering with this number of first PC's produces higher adjusted Rand indices than clustering with the original data. An entry marked "-" means that clustering with this number of first PC's produces worse clustering results. An entry marked "=" in the table means that the adjusted Rand indices are comparable. Since few entries in Table 4.2 are marked "+", we showed that clustering with the number of first

Table 4.2: Results of using the number of first PC's that cover at least 90% of the total variations in real data. The minimum number of first PC's that cover at least 90% of the total variations on each data is shown in brackets. An entry marked "+" means that using this number of first PC's produces higher adjusted Rand index than clustering the original data. An entry marked "-" means that the adjusted Rand index from this number of first PC's is lower than that from clustering with the original data. An entry marked "=" means that the adjusted Rand indices are comparable.

data (# PC's)	*CAST* *correlation*	*k-means* *correlation*	*k-means* *distance*	*average-link* *correlation*	*average-link* *distance*
ovary data (14)	-	+	-	+	-
cell cycle data (8)	=	-	-	-	+
sporulation (3)	-	NA	=	-	+

PC's that covers 90% of the variation in the data usually produces worse clustering results than clustering with the 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 ten replicates. We tested the null hypothesis that the adjusted Rand indices from clustering with the first PC's are comparable to that from the original data. The alternative hypothesis is that the adjusted Rand indices from clustering with the first PC's are greater than that from the original data. A low p-value suggests rejecting the null hypothesis. Table 4.3 shows the fraction of times for which the p-values lie below the 5% significance level. Hence, a small fraction means that clustering with the first PC's tend to produce relatively low adjusted Rand indices. In general, Table 4.3 shows that the adjusted Rand indices from the first components tend to be lower than those from the original synthetic data in most cases, especially when CAST with correlation, k-means with Euclidean distance and average-link with Euclidean distance are used. On the other hand, the adjusted Rand indices from the first components are often higher than those from the original data when k-means with correlation or average-link with correlation are used on the mixture of normal synthetic data and randomly resampled data. However, the latter results are not clear successes for PCA since (1) they assume that the

Table 4.3: Summary of results from the Wilcoxon signed rank test for the null hypothesis that the adjusted Rand indices from clustering with the first components are comparable to those from clustering with the original synthetic data sets. The alternative hypothesis is that the adjusted Rand indices from clustering with the first PC's are **greater** than those from clustering with the original data. Each entry shows the fraction of the number of times that the p-values from the Wilcoxon signed rank test lies below 5%.

synthetic data	*CAST correlation*	*k-means correlation*	*k-means distance*	*average-link correlation*	*average-link distance*
normal	0/22	20/22	0/21	15/22	1/21
resampled	0/22	13/22	2/22	21/22	4/22
cyclic	0/22	NA	0/21	0/22	6/21

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 p-values of k-means with correlation on the cyclic data are not available because the iterative k-means algorithm does not converge.

4.6 Conclusions

Our experiments on three 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 4.1 and Table 4.3). 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 those 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 for choosing the number of first PC's. Choosing PC's by the rule of thumb to cover 90% of the total variation in the data is usually not a good strategy (summarized in Table 4.2). 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 at least 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 usually include the first two PC's on most of our real and synthetic data sets (the sporulation data is the only exception). 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.

Our results applying the greedy and modified greedy approaches to the original data and random orthogonal bases show that clustering with PC's produces higher quality clusters than clustering with subsets of orthogonal bases and subsets of the original data when the number of components is small. This implies that there usually exist some "good" PC's that capture more cluster structure than subsets of random orthogonal bases and subsets of the original data. However, the quality of clusters produced by using the first few PC's are *not* higher than those produced by using subsets of random orthogonal bases. This suggests that the "good" PC's are usually not the first few PC's, and there is no well-established method to determine the "good" PC's without relying on the external criteria (which are seldom available in practice). When the number of components is large, there exist subsets of random orthogonal bases and subsets of the original data with comparable adjusted Rand indices to subsets of PC's (determined by the greedy and modified greedy approaches). This suggests that the relatively high adjusted Rand indices of large subsets of PC's determined by the greedy and modified greedy approaches may be an artifact of the search heuristic, and may not contain any more cluster information than random subsets of the data (either the original data or other transformed data).

Our empirical study shows that clustering with the PC's enhances cluster quality only when the right number 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.

Chapter 5

MODEL-BASED CLUSTERING AND DATA TRANSFORMATIONS FOR GENE EXPRESSION DATA

Many different heuristic-based clustering algorithms have been proposed to analyze gene expression data, for example, hierarchical algorithms [27], k-means [78], self-organizing maps [77] and many others. However, no clustering algorithm has emerged as the method of choice in the gene expression community. 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. Gaussian mixture models have been shown to be a powerful tool for clustering in many applications (for example, [5], [15], [58], [59]). 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. This provides a great advantage over heuristic clustering algorithms, for which there is no rigorous method to determine the number of clusters or the best clustering method.

5.1 Model-based clustering approach

5.1.1 The model-based framework

The mixture model assumes that each component (group) of the data is generated by an underlying probability distribution. Suppose the data $\mathbf{y}$ consist of independent multivariate observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$. Let G be the number of components in the data. The likelihood for the mixture model is

$$\mathcal{L}_{MIX}(\theta_1, \ldots, \theta_G|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i|\theta_k), \tag{5.1}$$

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 ($\tau_k \geq 0$ and $\sum_{k=1}^{G} \tau_k = 1$).

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

$$f_k(\mathbf{y}_i|\mu_k, \Sigma_k) = \frac{\exp\{-\frac{1}{2}(\mathbf{y}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{y}_i - \mu_k)\}}{\sqrt{\det(2\pi\Sigma_k)}}. \tag{5.2}$$

Geometric features (shape, volume, orientation) of each component k are determined by the covariance matrix Σ_k. Banfield and Raftery [5] proposed a general framework for exploiting the representation of the covariance matrix in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{5.3}$$

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 5.3 to vary results in a set of models within this general framework that is sufficiently flexible to accommodate data with widely varying characteristics. In particular, we consider five such models, outlined below. Constraining $D_k A_k D_k^T$ to be the identity matrix I corresponds to Gaussian mixtures in which each component is spherically symmetric. The *equal volume spherical* model (denoted 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 is a range of elliptical models with other constraints and fewer parameters. For example, with the parameterization $\Sigma_k = \lambda D A D^T$, each component is elliptical, but all have equal volume, shape and orientation (denoted EEE). All of these models are implemented in MCLUST [30]. Celeux and Govaert [16] 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 our experiments, we considered the equal volume spherical (EI), unequal volume spherical (VI), EEE and unconstrained (VVV) models as implemented in MCLUST [31], and the diagonal model as implemented by Murua *et al.*[63].

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 $1 \leq k \leq G$) 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. We initialized the EM algorithm with model-based hierarchical clustering ([23], [30]), but more often people initialize the EM algorithm with random starts or heuristically obtained partitions.

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 equal volume spherical model (EI), as computed by the EM algorithm [14]. 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 unequal volume spherical model (VI) 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 data set with different types of tissue samples. 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 trade-off 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 (experiments) 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. This is the topic of the next subsection.

5.1.2 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 trade-off 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 models with parameters θ_1 and θ_2 respectively. The *integrated likelihood* is defined as $p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$ where $k = 1, 2$ and $p(\theta_k|M_k)$ is the prior distribution of θ_k. The integrated likelihood represents the probability that data D is observed given that the underlying model is M_k. The Bayes factor [45] 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) [73]:

$$2 \log p(D|M_k) \approx 2 \log p(D|\widehat{\theta_k}, M_k) - \nu_k \log(n) = BIC_k \tag{5.4}$$

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 5.4, 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 [45].

Typically, different models of the model-based clustering algorithm are applied to a data set over a range of numbers of clusters. The BIC scores for the clustering results are computed for each of the models. The model and the number of clusters with the maximum BIC score are usually chosen for the data set.

5.2 Our Approach

Our main objective is to demonstrate the potential usefulness of the model-based approach with existing implementations. Since the model-based approach is based on the assumption that the data are distributed according to a mixture of Gaussian distributions, and we do not expect the raw gene expression data to satisfy the Gaussian mixture assumption, we explored the extent to which different transformations of gene expression data sets satisfy the Gaussian mixture assumption in Section 5.3. In order to show that the model-based approach finds relatively high quality clusters, we used both real and synthetic data sets with external criteria from Chapter 2 and compared clustering results to the external criteria using the adjusted Rand index [39] (see Section 2.4.1). In addition to applying different models of the model-based approach to data sets with external criteria, we also compared the quality of clustering results of CAST [10] to that of the model-based approach. In Chapter 3, we 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 equal volume spherical model (EI), we compared the quality of clusters obtained from the model-based approach to that of CAST using

correlation as the similarity metric.

5.2.1 Prior Work

We are aware of only two published papers attempting model-based formulations of gene expression clustering. Holmes and Bruno [37] formulate a model that appears to be equivalent to the unconstrained model defined above. Barash and Friedman [6] 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.

5.3 Data Transformations and the Gaussian mixture assumption

Before applying model-based clustering to gene expression data, we assessed the extent to which the Gaussian mixture assumption holds. Since we do not expect raw expression data to satisfy the Gaussian mixture assumption, we explored the degree of normality of each class of real expression data after various data transformations. In particular, we studied two types of data transformations: Box-Cox transformations and standardization.

The Box-Cox transformation [12] is a parametric family of transformations from y to $y^{(\lambda)}$ with parameter λ:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0. \end{cases} \tag{5.5}$$

The Box-Cox transformation subsumes many commonly used transformations, including the logarithm (log) transformation which is very popular for microarray data (for example, [76]). Standardization is another very popular data transformation step, for example, [77] and [78]. In standardization, the mean over all the experiments of a gene is subtracted from the expression level of the gene, and the difference is then divided by the standard deviation of the expression levels of the gene over all the experiments. Therefore, standardization captures the "patterns" of expression levels, and not the absolute expression levels.

5.3.1 Methodology to test Gaussian mixture assumption

In order to test the Gaussian mixture assumption, gene expression data sets with external criteria described in Chapter 2 were used. We tested the multivariate normality of *each class* in each data set. There are large collections of tests for multivariate normality. We used three different approaches: goodness of fit tests based on the empirical distribution function, e.g. Aitchison [2], skewness and kurtosis tests, e.g. Jobson [42], and maximum likelihood estimation of the transformation parameters, e.g. Andrews *et al.* [4].

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

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

$$\hat{\mu}_j^g = \sum_{i=1}^{n_g} x_{ij}/n_g, \tag{5.6}$$

$$\hat{\sigma}_{kj}^g = \sum_{i=1}^{n_g} (x_{ik} - \hat{\mu}_k^g)(x_{ij} - \hat{\mu}_j^g)/(n_g - 1). \tag{5.7}$$

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

Three different forms of empirical distribution functions (Anderson-Darling, Cramer-von Mises, and Watson) were used to measure departures of the sorted z_i^g values from the order statistics of the uniform distribution. Assuming that z_i^g are the sorted values from class g, the Anderson-Darling statistic is defined as $Q_A = -\{\sum_{i=1}^{n_g}(2i-1)\{\log z_i^g + \log(1 -$

$z^g_{n_g+1-i})\} - n_g\}/n_g$. The Cramer-von Mises statistic is defined as $Q_C = \sum_{i=1}^{n_g}\{z_i^g - (2i - 1)/(2n_g)\}^2 + 1/(12n_g)$. The Watson statistic is defined as $Q_W = Q_C - n_g(\bar{z} - \frac{1}{2})^2$ where $\bar{z} = \sum_{i=1}^{n_g} z_i^g/n_g$. Critical values of the empirical distribution function test statistics are given in Aitchison [2]. We used the critical values corresponding to the 1% significance level. For each class, we computed the empirical distribution function test statistics for each of the Anderson-Darling, Cramer-von Mises, and Watson forms using the z_i^g's. If a given test statistic for experiment j is greater than the critical value, we say that the marginal distribution of experiment j shows departure from normality.

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

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

Skewness and Kurtosis: Skewness measures the amount of asymmetry in a distribution. For a normal distribution, the skewness is 0. Kurtosis measures the extent to which the data are peaked or flat relative to the normal distribution. For the standard normal distribution, the kurtosis is 3. We computed the skewness and kurtosis of each class g in the data. Let $m_{ir} = (\mathbf{x_i} - \hat{\mu}^{\mathbf{g}})^{\mathbf{T}}(\hat{\mathbf{\Sigma}}^{\mathbf{g}})^{-\mathbf{1}}(\mathbf{x_r} - \hat{\mu}^{\mathbf{g}})$, where $i, r = 1, \ldots, n_g$. Multivariate skewness and kurtosis are defined by $\sum_{i=1}^{n_g}\sum_{r=1}^{n_g} m_{ir}^3/n_g^2$ and $\sum_{i=1}^{n_g} u_i^2/n_g$, and there are distributions for both the multivariate skewness and kurtosis [56]. A small p-value suggests the multivariate normal

assumption to be questionable.

Maximum likelihood estimation of the transformation parameters: The parameter λ in the Box-Cox transformation in Equation 5.5 is estimated by maximum likelihood using the observations [4]. The estimated value of λ suggests the scale on which the data are closest to normality, and the extent to which the data on other scales deviate from normality.

5.3.2 Results of testing the Gaussian mixture assumption

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

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

Ovary data: Table 5.1a shows the results of the Aitchison tests on each of the four classes in the ovary data. In the marginal test, if the test statistics of an experiment j from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level, we adopt the shorthand convention of saying that experiment j *violates* the normality assumption. The column **m** in Table 5.1a shows the number of violations from the 24 marginal tests on each class in the ovary data. Similarly, the column **b** in Table 5.1a shows the number of violations from $\binom{24}{2} = 276$ bivariate angle tests on each class in the ovary data. The column **r** has an entry 1 if the test statistics from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level in the radius test. Otherwise, the column **r** has an entry 0. The results from Table 5.1a suggest that the square root transformation is closer to multivariate normal

Table 5.1a: Results of Aitchison tests on the ovary data. Column **m** shows the number of violations from the marginal tests. Column **b** shows the number of violations from the bivariate angle tests. Column **r** has an entry 1 if the radius test shows violation of the normality assumption. A "-" entry means that the corresponding test is not available.

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

Table 5.1b: p-values of skewness and kurtosis on the ovary data.

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

than the log transformation. On the square root transformed data, the marginal test shows that only one experiment (out of 24) deviates from normality in class 1. Similarly, class 2 has 6 experiments, class 3 has 4 experiments and class 4 has 3 experiments that deviate from marginal normality. None of the classes in the square root transformed data shows any deviation in the bivariate angle or radius tests. On the standardized data, the radius tests are not applicable, so the **r** columns for the standardized data are marked "-" in Table 5.1a.

Table 5.1b shows the p-values of skewness and kurtosis for each class on the raw, log and square root transformed ovary data. Small p-values indicate deviations from the skewness and kurtosis criteria. From Table 5.1b, class 2 deviates from the skewness and kurtosis criteria in the raw, log and square root transformed data. On the other hand, class 4

does not violate the skewness or kurtosis criteria in any of the data transformations. Both the square root and log transformations improve skewness in the raw data, but the log transformation makes class 1 skewed. To summarize, the skewness and kurtosis tests show the same overall picture as the Aitchison tests: the square root transformation is relatively close to multivariate normal.

Table 5.1c: Estimates of the transformation parameters and log-likelihood values for the ovary data.

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

Table 5.1c shows the results of the maximum likelihood estimation of the transformation parameters on each of the four classes on the raw ovary data. $\mathcal{L}_{max}(0.5)$ and $\mathcal{L}_{max}(0)$ are the log-likelihoods of the square root and log transformations respectively. From Table 5.1c, the optimal parameters for the Box-Cox transformation ($\hat{\lambda}$) lie between 0.40 and 0.73 for the four classes in the ovary data. Comparing the log-likelihood values of the square root transformation to those of the log transformation shows that the square root transformation is closer to the multivariate normal distribution in all four classes.

Yeast cell cycle data with the 5-phase criterion: Table 5.2a shows the results of the Aitchison tests on the yeast cell cycle data with the 5-phase criterion. The results from Table 5.2a show that the log transformed yeast cell cycle data is relatively close to the multivariate normal distribution than the square root transformation. With the log transformation, classes 1, 3, and 4 show no deviation from any of the marginal, bivariate angle and radius tests. The only deviations from normality in this data set are: class 2 shows deviation from the radius test, and one experiment (out of 17) in class 5 shows deviation from marginal normality. The Aitchison tests show that the log transformation greatly enhances normality in all of the 5 classes: the raw data shows significant deviations

Table 5.2a: Results of Aitchison tests on the yeast cell cycle data with the 5-phase criterion. Abbreviations have the same meanings as in Table 5.1.

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

Table 5.2b: p-values of skewness and kurtosis on the yeast cell cycle data with the 5-phase criterion.

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

from the marginal, bivariate angle and radius tests in all of the 5 classes. The standardized yeast cell cycle data is also much more Gaussian than the raw data, but not as much as the log transformed data.

Table 5.2b portrays a different picture than the Aitchison tests: the raw, square root and log transformed data all show deviations from the skewness and kurtosis criteria. However, the log transformation seems to show relatively less deviation.

Table 5.2c supports the conclusions from the other approaches: the log transformation has higher log-likelihoods than the square root transformation. The estimates $\hat{\lambda}$ are between 0.14 and 0.22 for all 5 classes.

Yeast cell cycle data with the MIPS criterion: In general, the Aitchison tests, the

Table 5.2c: Estimates of the transformation parameters and log-likelihood values for the yeast cell cycle data with the 5-phase criterion.

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

Table 5.3a: Results of Aitchison tests on the yeast cell cycle data with the MIPS criterion. Abbreviations have the same meanings as in Table 5.1.

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

skewness and kurtosis tests, and the maximum likelihood estimation all show similar patterns to the 5-phase criterion: the log transform is relatively more Gaussian than the square root transformation (see Tables 5.3a, b and c). However, class 4 (ribosomal proteins) shows significantly more deviations from normality with very low p-values for both the skewness and kurtosis tests using the log and square root transformations.

Synthetic data: We also tested the Gaussian mixture assumption on the synthetic data sets. As expected, the mixture of normal distributions based on the ovary data closely follows the Gaussian mixture assumption. Table 5.4 shows the number of violations of the Aitchison tests for each of the ten replicates of the randomly resampled ovary data. The randomly resampled ovary data shows some significant deviations from normality, especially class 2. (Class 2 also shows significant deviations from normality in the original ovary data.)

Table 5.3b: p-values of skewness and kurtosis on the yeast cell cycle data with the MIPS criterion.

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

Table 5.3c: Estimates of the transformation parameters and log-likelihood values for the yeast cell cycle data with the MIPS criterion.

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

Since the randomly resampled ovary data is generated from the standardized ovary data, the results of the radius tests are not available. We also tested the normality of the cyclic data, and the classes from the cyclic data show significant deviations from the Gaussian assumption (results not shown here).

5.4 Results of applying model-based clustering

In this section, we show the results of applying model-based clustering to both synthetic and real gene expression data. We applied different models from the model-based approach and CAST (a leading heuristic-based clustering algorithm) to each data set over a range of different numbers of clusters. The clustering results are compared to the external criteria with the adjusted Rand index. The BIC scores for the clustering results over a range of

Table 5.4: Results of Aitchison tests on the randomly resampled ovary data. Abbreviations have the same meanings as in Table 5.1.

Replicate	*class 1*			*class 2*			*class 3*			*class 4*		
	m	b	r	m	b	r	m	b	r	m	b	r
1	4	0	-	8	4	-	7	0	-	8	0	-
2	7	0	-	10	2	-	9	0	-	3	0	-
3	5	0	-	9	7	-	8	0	-	8	0	-
4	4	0	-	9	4	-	8	0	-	7	0	-
5	4	0	-	10	1	-	10	17	-	7	3	-
6	5	0	-	6	12	-	8	0	-	6	0	-
7	6	0	-	11	13	-	8	0	-	6	0	-
8	6	0	-	9	2	-	10	0	-	7	2	-
9	5	0	-	10	0	-	9	1	-	6	0	-
10	9	0	-	9	12	-	9	0	-	6	0	-

numbers of clusters are also computed. Hence, two graphs are typically shown for each data set: the adjusted Rand index against the number of clusters, and the BIC score against the number of clusters.

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 unconstrained model (VVV), 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 the number of parameters to be estimated for the unconstrained model (VVV) is 324, which is greater than the number of observations (235) in the data set. In order to compute the BIC scores for the unconstrained model (VVV), we generated the mixture of normal distributions based on the ovary data with 2350 observations, such that the size of each of the four classes is ten times that of the original ovary data. Even with 2350 observations, when the number of clusters is greater than 7, the number of parameters to be estimated for the unconstrained

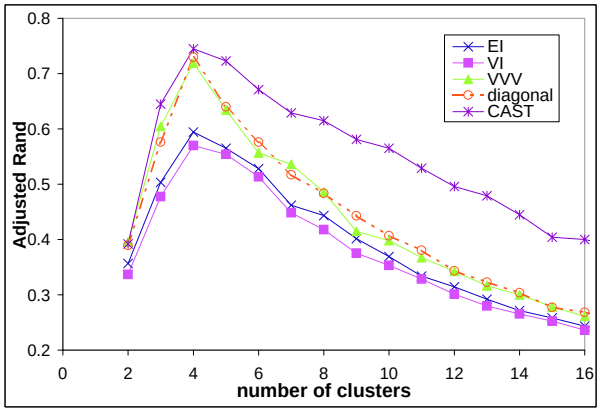

Figure 5.1a: Average adjusted Rand indices for the mixture of normal synthetic data.

model (VVV) 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 (see Section 2.1.3), the algorithm may not produce a result for every number of clusters. Hence, in the following result graphs, not all data points are available for CAST.

5.4.1 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.

Mixture of normal distributions based on the ovary data:

Figure 5.1a 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 results for the EEE model are not available because of the long running time of the hierarchical step for this large synthetic data set.) The average adjusted Rand indices reach the maximum at 4 clusters, with the unconstrained model (VVV), 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 unconstrained model (VVV) fits the data the best.

Figure 5.1b 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

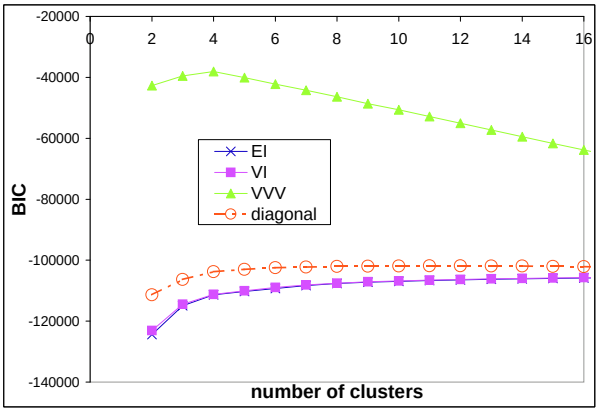

Figure 5.1b: Average BIC scores for the mixture of normal synthetic data.

achieved by the unconstrained model (VVV) at 4 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.

Randomly resampled ovary data:

Figure 5.2a 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 5.2b 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.

Our results on testing the Gaussian mixture assumption (Section 5.3.2) showed that the randomly resampled ovary data exhibit significant deviations from the normality assumption. However, the diagonal model still produces higher quality clusters than CAST. This shows that the model-based clustering approach not only produces high quality clusters when the Gaussian mixture assumption is satisfied (*i.e.*, the mixture of normal distributions based on the ovary data), but can also do so when the assumption is violated (*i.e.*,

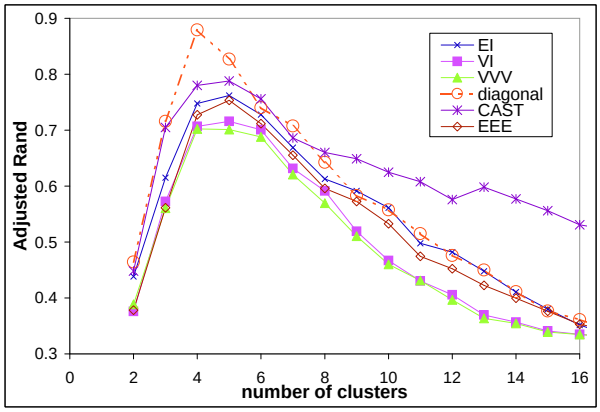

Figure 5.2a: Average adjusted Rand indices for the randomly resampled ovary data.

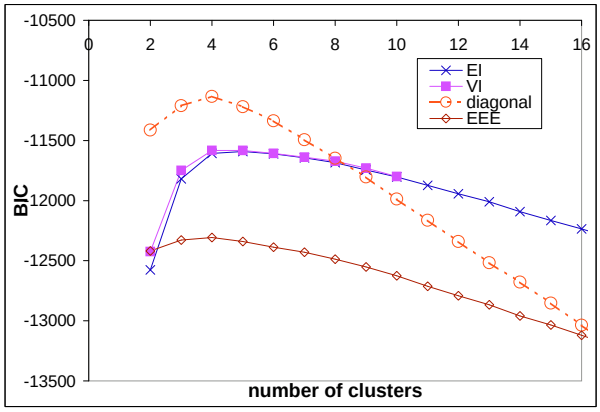

Figure 5.2b: Average BIC scores for the randomly resampled ovary data.

the randomly resampled ovary data).

Cyclic data:

Figure 5.3a shows that the average adjusted Rand indices of CAST and five different models from the model-based approach are comparable. This synthetic data set contains ten classes. The adjusted Rand indices from CAST are higher than any of the model-based approaches at 10 clusters. In practice, however, one would not know the correct number of clusters, so its performance at the number of clusters that one would select is the most relevant. Furthermore, all of the algorithms show average adjusted Rand indices peaking around 6 or 7 clusters. This set of synthetic data consists of classes with varying sizes, with some very small classes, which can be problematic for most clustering methods

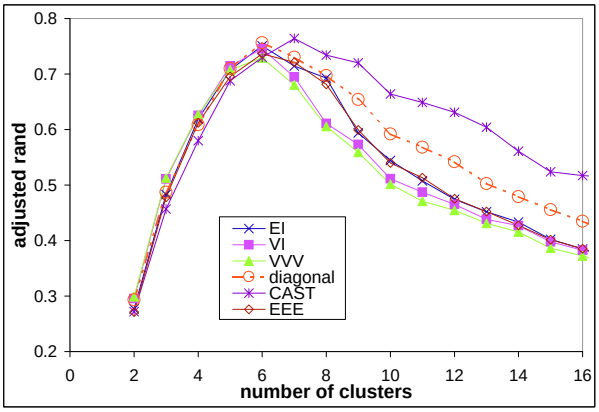

Figure 5.3a: Average adjusted Rand indices for the cyclic data.

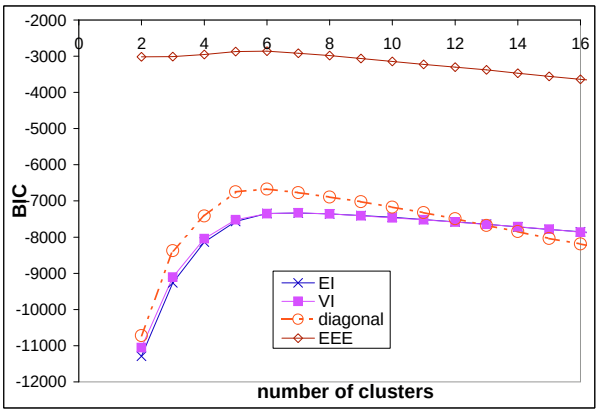

Figure 5.3b: Average BIC scores for the cyclic data.

including the model-based approach (small clusters make estimation of parameters difficult). In Figure 5.3b, the BIC scores of the models also peak around 6 to 7 clusters, with the EEE model showing higher BIC scores (there are too few data points to compute BIC scores for the unconstrained model). Our results show that the BIC scores select the number of clusters that maximizes the adjusted Rand indices, and the quality of clusters are comparable to CAST at 6 or 7 clusters.

5.4.2 Gene expression data sets

Ovary data:

Figure 5.4a shows that the spherical models (EI and VI) and the EEE model produce

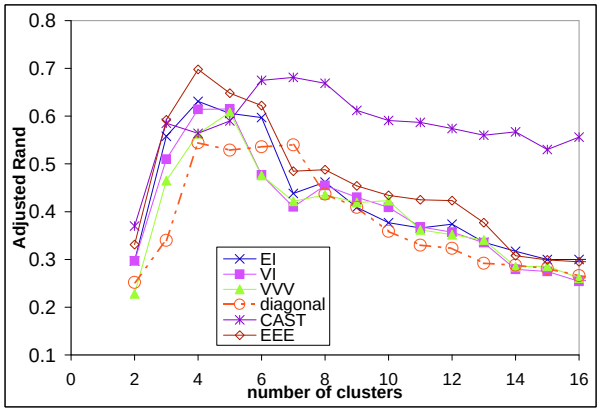

Figure 5.4a: Adjusted Rand indices for the square root transformed ovary data.

higher quality clusters than CAST and the diagonal and unconstrained models (VVV) at 4 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. This is because the heuristic-based CAST algorithm tends to separate the outliers while the model-based approach tends to split a cluster into small clusters when the number of clusters is large. In Figure 5.4b, the EEE model has its first local maximum BIC score at 4 clusters (the correct number of classes), the diagonal model has its global maximum BIC score at 4 clusters, and the BIC curves of the spherical models (EI and VI) show a bend at 4 clusters. However, the spherical models (EI and VI) at 8 clusters achieve the highest BIC scores. Close inspection of the data reveals that the 8 cluster solution selected by BIC analysis is still a meaningful clustering — it differs from the external criterion mainly in that the larger classes have been split into 2 or 3 clusters (which may reflect differences in the constituent cDNAs, for example). Even though real expression data may not fully reflect the class structure due to noise, the BIC analysis gives a reasonable hint to the number of clusters.

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 4 clusters (figures not shown). On the standardized ovary data, the adjusted Rand indices of clusters produced by EEE and EI are comparable

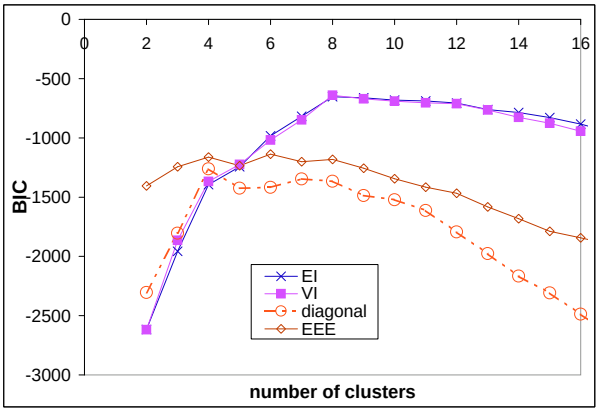

Figure 5.4b: BIC scores for the square root transformed ovary data.

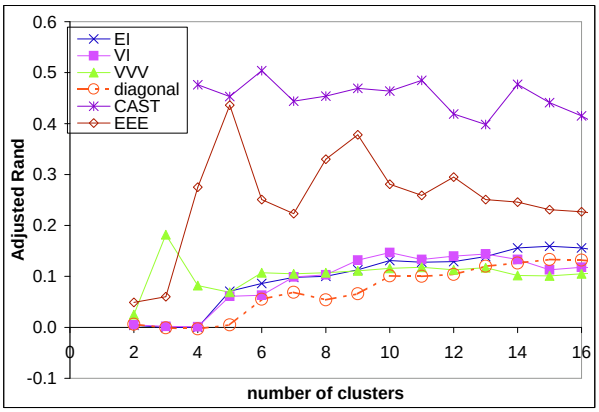

Figure 5.5a: Adjusted Rand indices for the log transformed yeast cell cycle data with the 5-phase criterion.

to that from CAST. The BIC curves start to flatten at around 4 clusters on the standardized ovary data, but the maximum occurs at around 7 clusters.

Yeast cell cycle data with the 5-phase criterion:

With the exception of the EEE model, all the other models show considerably lower adjusted Rand indices than those from CAST on the log transformed yeast cell cycle data with the 5-phase criterion (Figure 5.5a). Figure 5.5b shows that the BIC analysis selects the EEE model at 5 clusters, which is the number of classes in this data. The model selected by the BIC analysis (EEE) at the selected number of clusters (5) produces similar adjusted Rand index to CAST at the same number of clusters.

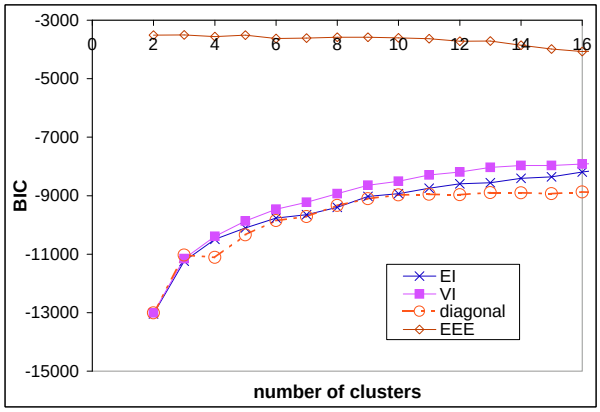

Figure 5.5b: BIC scores for the log transformed yeast cell cycle data with the 5-phase criterion.

The standardized yeast cell cycle data set shows a very different picture from the log transformed data (Figure 5.6): the equal volume spherical model (EI) achieves higher adjusted Rand indices than CAST at 5 clusters. A careful study of the nature of the data shows that this is no surprise. The yeast cell cycle data set consists of time course data, and so all 17 experiments are highly correlated (unlike the ovary data). Figure 5.7a shows a pairs plot of the first four time points of the log transformed yeast cell cycle data. Data points from each of the five classes are represented by different symbols. The pairs plots of the remaining 13 time points show a similar pattern. Figure 5.7a 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 cluster 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. Figure 5.7b is a pairs plot of the first four time points of the standardized yeast cell cycle data. 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 equal volume spherical model (EI) captures the class information on the standardized data. The BIC analysis (figure not shown) selects model EEE at 5 clusters. In the case of CAST, correlation coefficient is used as the similarity measure, and correlation captures the "general patterns" across experiments even when

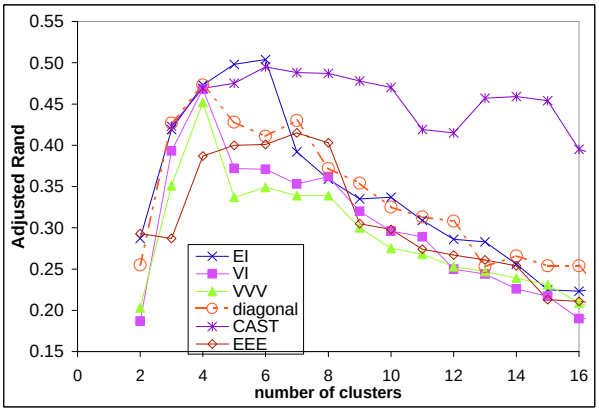

Figure 5.6: Adjusted Rand indices for the standardized yeast cell cycle data with the 5-phase criterion.

the data set is log-transformed (without standardization).

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.*, $\log(x_{ij}/\sum_{k=1}^{17} x_{ik})$, 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 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.

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 5-phase criterion: CAST produces much higher quality clusters than the model-based approach on the 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 5-phase criterion, the equal volume spherical (EI) model produces comparable adjusted Rand indices to CAST (Figure 5.8a) on the standardized data. The BIC curve of model EI shows a bend at 4 clusters, which is the number of classes

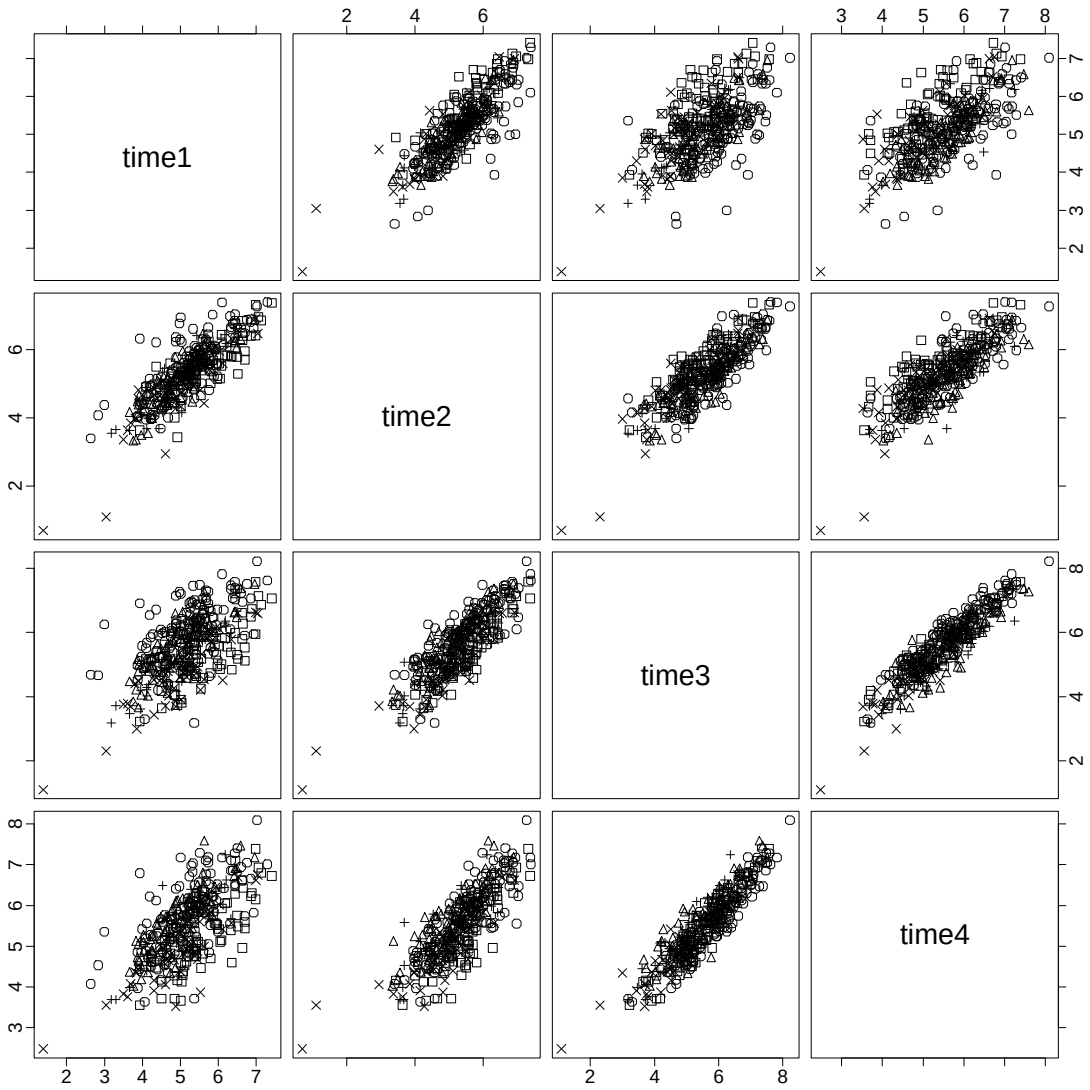

Figure 5.7a: Visualization of the log transformed yeast cell cycle data with the 5-phase criterion.

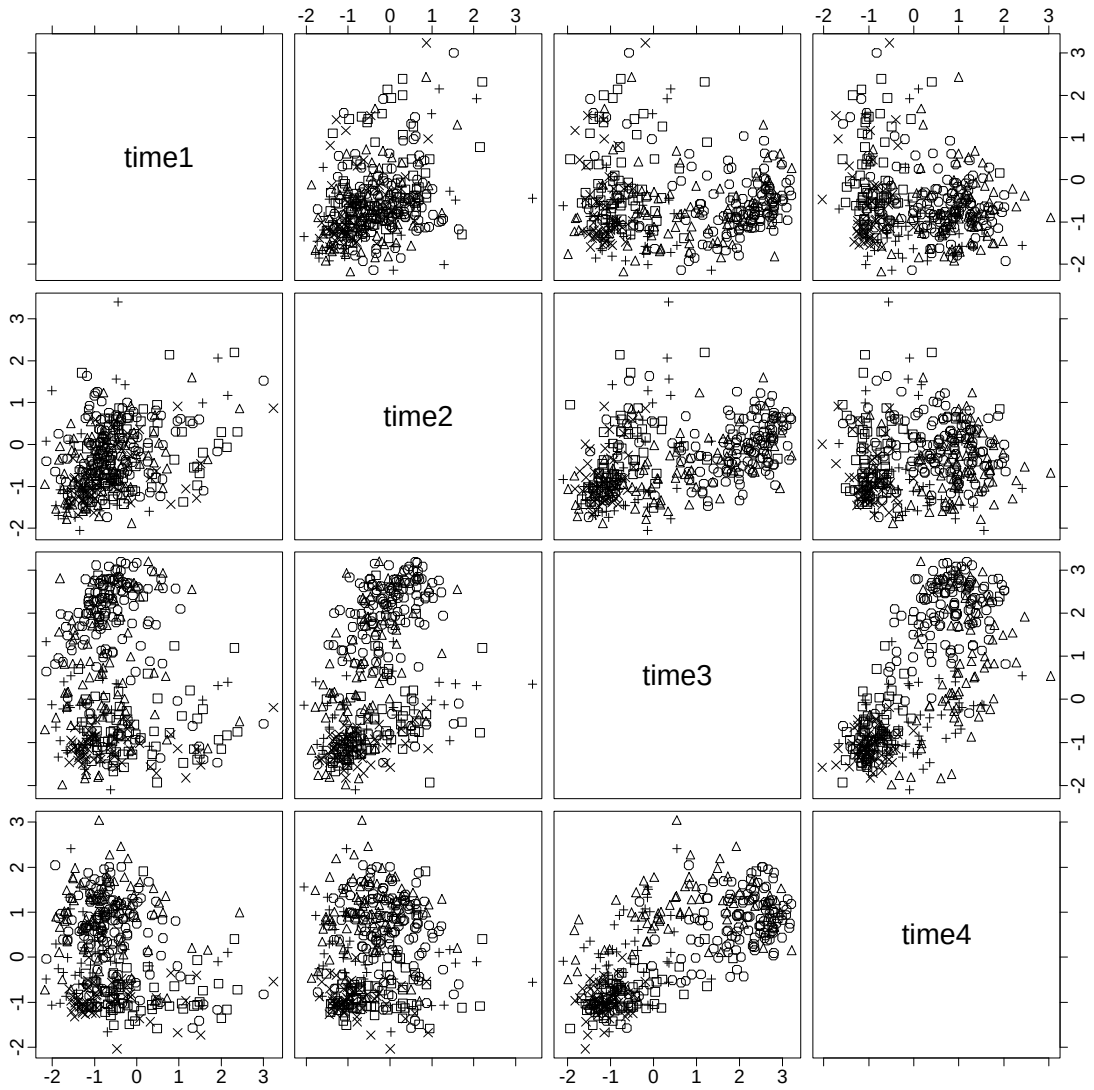

Figure 5.7b: Visualization of the standardized yeast cell cycle data with the 5-phase criterion.

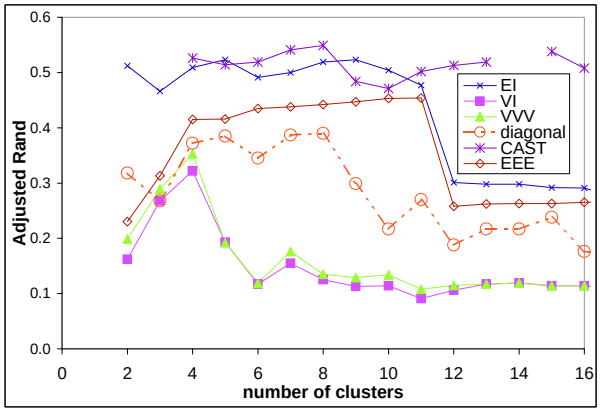

Figure 5.8a: Adjusted Rand indices for the standardized yeast cell cycle data with the MIPS criterion.

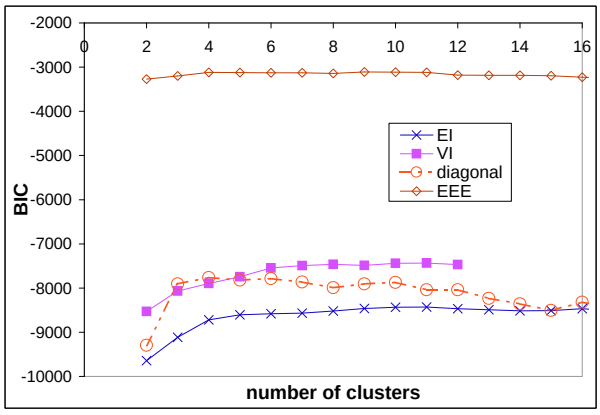

Figure 5.8b: BIC scores for the standardized yeast cell cycle data with the MIPS criterion.

in this data (Figure 5.8b). However, the BIC analysis selects the EEE model at 4 clusters. Note that although the BIC analysis does not select the best model, it does select the 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.

5.4.3 Summary of Results

With our synthetic data sets, the model-based approach not only produced higher adjusted Rand indices, 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 unconstrained

model (VVV) produced the highest quality clusters and the BIC analysis chose the right model and the number of clusters. On the randomly resampled 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 randomly resampled data showed considerable deviation from the Gaussian mixture assumption. On the cyclic data sets (which showed significant deviations from the Gaussian mixture assumption and contained very small classes), we showed that the model-based approach and CAST (a leading heuristic-based approach) produced comparable quality clusters, and the BIC analysis selected the number of clusters that maximized the average adjusted Rand index.

We also showed the practicality of the model-based approach on real gene expression data sets. On the ovary data, the model-based approach achieved slightly better results than CAST, and the BIC analysis gave a reasonable indication of the number of clusters in the transformed data. On two different subsets of the yeast cell cycle data with different external criteria, the equal volume spherical model (EI) and EEE model produced comparable results to CAST on the standardized data. The BIC scores from the EEE model were maximized at the correct number of clusters. The results are summarized in Table 5.5.

5.4.4 Effect of initialization methods

The EM algorithm of the model-based approach is initialized with a model-based hierarchical algorithm. In agglomerative hierarchical clustering methods, each object initially starts in its own cluster, and then pairs of clusters are successively merged until the desired number of clusters is reached. (In some applications, agglomerative hierarchical clustering methods may start with groups of objects in order to speed up the algorithms and to reduce memory requirements.) Different criteria for choosing the pair of clusters to be merged lead to different types of hierarchical clustering algorithms (see Chapter 2). In the case of model-based hierarchical clustering, a maximum-likelihood pair of clusters is chosen for merging in each step. However, the maximum likelihood of a pair of clusters depends on the model assumed. Hence, different model assumptions in the hierarchical initialization step can

Table 5.5: Summary of results on real expression data. 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.

data (# classes)	*transf*	*max adj Rand*	*BIC analysis*	
		at # classes	model selected	notes
ovary (4)	log	EEE	VI (9)	EI,VI,diag: bend at 4
ovary (4)	sqrt	EEE	VI (8)	EEE: local max at 4
ovary (4)	std	EI	EEE (7)	EI,VI, EEE: bend at 4
5-phase cell cycle (5)	log	EEE, CAST	EEE (5)	
5-phase cell cycle (5)	std	EI	EEE (5)	
MIPS cell cycle (4)	log	CAST	EEE (4)	
MIPS cell cycle (4)	std	CAST, EI	EEE (4)	

potentially lead to different clustering results.

We studied the effect of different initialization methods on clustering results in terms of both the BIC scores and the adjusted Rand indices. In the MCLUST implementation [31], the default initialization method is the model-based hierarchical algorithm using the unconstrained model (VVV). The results shown in Section 5.4.1 and Section 5.4.2 are generated using the same model in the hierarchical and the EM steps. For example, the results for the equal volume spherical model (EI) are generated using the EI model in both the hierarchical and the EM steps. We compared the clustering results using the unconstrained model (VVV) in the hierarchical initialization step to clustering results using the same model in the initialization step as the EM algorithm.[1] In addition, since the EM algorithm may converge to a local optimal solution instead of a global optimum, we studied the quality of clustering results initialized with the true classes. Table 5.6 shows the adjusted Rand indices and BIC scores at the number of classes for the ovary data and the yeast cell cycle data with the 5-phase criterion for three initialization methods (same model as the EM step, VVV

[1]The model-based hierarchical clustering step using the EEE model requires considerably more running time than the other models [29].

initialization and initialization with the classes). Table 5.6 also shows the average adjusted Rand indices and average BIC scores (over the ten replicates) at the number of classes for the randomly resampled ovary data. The results on the yeast cell cycle data with the MIPS criterion and the mixture of normal distributions based on the ovary data are similar, and hence not shown. We ran experiments over a range of numbers of clusters on both real and synthetic data sets, but only selected results are shown in Table 5.6.

We observed that different initialization methods (even initialization with the true classes) yield comparable adjusted Rand indices and comparable BIC scores for the spherical models (EI and VI) on both real and synthetic data sets. In fact, among the results from the spherical models on different data transformations or different data sets shown in Table 5.6 (a total of 12), five such data transformations or data sets produce exactly the same clustering results for all three initialization methods. On the contrary, none of the results for the elliptical models (EEE and VVV) produce the same clustering results for all three initialization methods on any data set or any data transformation. In fact, different initialization methods produce clustering results with varying quality for the elliptical models. For example, the EEE model with the unconstrained model (VVV) initialization produces much lower adjusted Rand indices (0.014) and much lower BIC scores (-3791) than the EEE model with EEE initialization (adjusted Rand index = 0.436 and BIC score = -3514) on the log transformed yeast cell cycle data with the 5-phase criterion (see Table 5.6). Moreover, initialization with the true classes produces relatively higher adjusted Rand indices for the elliptical models (EEE and VVV). For example, on the standardized ovary data, and the standardized yeast cell cycle data with both criteria, the EEE and VVV models with initialization with the true classes produce clustering results with perfect adjusted Rand indices. In general, when the BIC scores from the VVV initialization and from the same model initialization differ significantly, the BIC score tends to favor the initialization method that produces a higher adjusted Rand index. However, unlike initialization with the unconstrained model (VVV), initialization with the true classes sometimes produces a lower BIC scores than other initialization methods.

In addition to initializations with the same model as in the EM step, initializations with the unconstrained model (VVV), and initializations with the true classes, we also studied

Table 5.6: Selected results of the effect of initializations on real and synthetic expression data sets. Three different types of initialization methods are shown: same initialization as the EM step, initialization with the unconstrained model (VVV), and initialization with the true classes. The adjusted Rand indices and BIC scores at the number of classes are shown for all three initialization methods. For the randomly resampled ovary synthetic data, the average adjusted Rand indices and average BIC scores over the ten replicates are shown. The "-" entries mean that the BIC scores are not available from MCLUST.

data	*transf*	*init*	*EI*		*VI*		*EEE*		*VVV*	
(# classes)			*Rand*	*BIC*	*Rand*	*BIC*	*Rand*	*BIC*	*Rand*	*BIC*
ovary	log	same	0.622	-6723	0.528	-6413	0.687	-6770	0.667	-9636
(4)		VVV	0.622	-6723	0.528	-6413	0.682	-6781	0.667	-9636
		true	0.622	-6723	0.550	-6524	0.701	-6736	0.850	-9930
ovary	sqrt	same	0.631	-1393	0.614	-1368	0.698	-1162	0.563	-
(4)		VVV	0.644	-1398	0.614	-1368	0.648	-1172	0.563	-
		true	0.644	-1398	0.648	-1393	0.693	-1156	0.872	-4727
ovary	std	same	0.670	-10729	0.555	-10407	0.661	-5933	0.490	-8785
(4)		VVV	0.670	-10729	0.555	-10407	0.490	-6044	0.490	-8785
		true	0.670	-10729	0.555	-10407	1.000	-6244	1.000	-9544
cell cycle	log	same	0.071	-10105	0.061	-9864	0.436	-3514	0.069	-5624
5-phase		VVV	0.071	-10105	0.061	-9864	0.014	-3791	0.069	-5624
(5)		true	0.071	-10105	0.075	-10007	0.464	-3514	0.371	-5452
cell cycle	std	same	0.498	-13138	0.372	-12958	0.400	-3181	0.337	-4982
5-phase		VVV	0.367	-13244	0.372	-12958	0.337	-3225	0.337	-4982
(5)		true	0.498	-13138	0.480	-12993	1.000	-3465	1.000	-5531
re-	none	same	0.748	-11606	0.707	-11583	0.728	-12306	0.702	-
sampled		VVV	0.753	-11607	0.707	-11583	0.698	-12302	0.702	-
(4)		true	0.753	-11607	0.707	-11583	0.993	-12325	0.800	-15715

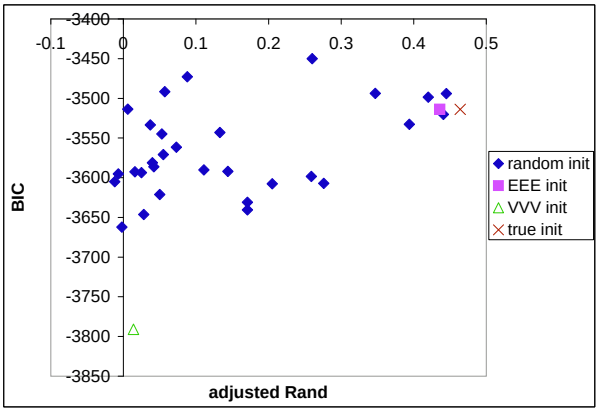

Figure 5.9: Plotting the BIC scores and adjusted Rand indices for the EEE model with random initializations on the log transformed yeast cell cycle data with the 5-phase criterion.

the adjusted Rand indices and BIC scores with random initializations. Figure 5.9 plots the BIC scores against the adjusted Rand indices for the EEE model with 30 random initializations on the log transformed yeast cell cycle data with the 5-phase criterion. Figure 5.9 shows that clustering results with high adjusted Rand indices tend to have high BIC scores. The hierarchical initialization with the EEE model and initialization with the true classes produce relatively high BIC scores and adjusted Rand indices. Out of the 30 random initializations, there are three clustering results with adjusted Rand indices above 0.4. In practice, different models need to be run over a range of numbers of clusters to determine the model and the number of clusters for the data. The hierarchical initialization step needs to be run only once before the EM step for different numbers of clusters. Hence, it is more computationally efficient to use hierarchical initializations than many trials of random initializations. Furthermore, the quality of clustering results from hierarchical initialization methods are relatively high. (The clustering result obtained from the EEE model with EEE initialization gives an adjusted Rand index of 0.436 on the log transformed yeast cell cycle data.) Therefore, we recommend against random initializations, and suggest initialization with one of the model-based hierarchical methods, and then choosing a good hierarchical initialization method using the BIC scores.

5.5 Conclusions and Future Work

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 tends to produce higher quality clusters than a leading heuristic-based algorithm with the key advantage of suggesting the numbers of clusters. It is interesting to note that simple models, like the equal volume spherical model (EI) and the elliptical EEE model, produced relatively high quality clusters on all of our transformed data sets. The EEE model even determined the right number of clusters on two different subsets from the yeast cell cycle data set with different external criteria. On the ovary data set, the BIC 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.

We also showed that the spherical models (EI and VI) are not very sensitive to the model used in the hierarchical initialization step, while the elliptical models (EEE and VVV) are more sensitive to initialization methods. We recommend using different initialization methods for the elliptical models, and using the BIC scores to determine the best models for the hierarchical initialization step.

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 dissertation, 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 ([23], [30], [70]).

Chapter 6

CASE STUDY: BARRETT'S ESOPHAGUS

The Barrett's esophagus data set provides us with a test bed for some of our analytical techniques. In this chapter, we will describe the experimental details and results of computational analyses for the Barrett's esophagus data.

6.1 Introduction

Barrett's esophagus is a pre-cancer condition in which the normal squamous epithelium in the esophagus is replaced by Barrett's epithelium. Barrett's metaplasia develops as a complication in 10-20% of patients with chronic gastroesophageal reflux disease [34], [65]. Since the mid 1970s, the incidence of Barrett's-associated adenocarcinoma has increased more rapidly than that of any other cancer in the United States [11]. Patients with Barrett's esophagus typically have symptoms of gastroesophageal reflux, such as heartburn or indigestion and they frequently seek medical attention before they develop cancer. Barrett's epithelium can be safely visualized and biopsied during upper gastrointestinal endoscopy. At the present time, total removal of Barrett's epithelium requires esophagectomy, a procedure with substantial morbidity and mortality. However, a systematic protocol of endoscopic biopsies can detect early curable cancers arising in Barrett's esophagus. Therefore, the standard of care for many patients includes endoscopic biopsy surveillance for the early detection of cancer.

The development of Barrett's metaplasia is fundamentally related to tissue differentiation. The phenotype of Barrett's metaplasia has been described by histologic, electron microscopic, immunohistochemical and biochemical studies, and the results show a surprisingly complex epithelium that shares features with duodenal, gastric and squamous esophageal epithelia. By electron microscopy, Barrett's metaplasia resembles small intestine [52], [51]. Barrett's metaplasia also has some features in common with gastric mucosa,

including mucus secretory capacity and mucus granules [52]. Barrett's metaplasia also shares some features with squamous esophageal cells, including expression of both squamous and columnar cytokeratins [69]. Recent microarray studies have shown that cancers, although highly variable, can be categorized into different classes based on the presence of distinctive expression signatures (reviewed in Young [82]). However, little is known about the molecular phenotype of human metaplasia *in vivo*. The ability to sample Barrett's epithelium and the surrounding normal tissues provides a unique *in vivo* human model to use microarray technology to compare a premalignant metaplastic tissue with the surrounding normal upper GI tissues, including squamous, gastric and duodenal epithelia. Moreover, identifications of gene clusters with relatively higher or lower expression levels in the premalignant tissue than the surrounding normal upper GI tissues may shed light on genes associated with Barrett's esophagus.

6.2 Experimental Details and Data Pre-processing

Endoscopic biopsies from Barrett's epithelium (BE), squamous epithelium (Sq), gastric epithelium (GAS) and duodenal epithelium (DUO) were collected from a series of patients during routine surveillance. Endoscopic biopsies of each tissue were pooled from two to four patients for each experiment. We collected sufficient material for 4 pools each of Barrett's epithelium and of esophageal squamous epithelium, and 3 pools each of gastric and duodenal biopsies. Poly A^+ RNA was prepared from the pooled tissue samples, and each poly A^+ sample was used to prepare double-stranded cDNA with a T7 promoter. Subsequently, fluorescently labeled cRNA, generated by in vitro transcription (IVT) of the cDNA template, was used to interrogate Affymetrix HU6800 and FL6800 chips. Please refer to our paper [8] for more experimental details.

In our first set of experiments, four pools of Barrett's epithelium, four pools of squamous epithelium, one pool of gastric epithelium and one pool of duodenal epithelium were used to interrogate the Affymetrix Hu6800 chips (a total of ten experiments). Let us denote the first set of experiments {BE1, BE2, BE3, BE4, Sq1, Sq2, Sq3, Sq4, GAS1, DUO1}. In our second set of experiments, one pool of Barrett's epithelium, one pool of squamous

epithelium, two pools of duodenal epithelium and two pools of gastric epithelium were used to interrogate the Affymetrix FL6800 chips (a total of six experiments). Let us denote the second set of experiments {BE5, Sq5, GAS2, GAS3, DUO2, DUO3}. The pools of Barrett's epithelium and squamous epithelium used in the second set of experiments (the FL6800 chips) were identical to one of the four pools used in the first set of experiments (the Hu6800 chips). In particular, BE4 and BE5 were derived from the same pool of tissue samples of Barrett's epithelium, and Sq2 and Sq5 were derived from the same pool of tissue samples of squamous epithelium. Note that each of the two sets of experiments cover all four types of tissue samples.

The Affymetrix Hu6800 and FL6800 chips cover approximately 7000 genes. The two types of chips contain the same genes. However, the Hu6800 format divides the 7000 genes into four separate physical chips (namely, A,B,C,D), while the FL6800 format has all the 7000 genes on one physical chip. Figure 6.1 is a cartoon of the data set. In the first set of experiments, approximately one quarter of the 7070 genes are on each of the A,B, C, D chips, and the A, B, C, D chips contain the same genes across different experiments. From our experience, the four chips in the Hu6800 format can have very different overall intensities. For example, in experiment E_1, the A chip is much brighter than the D chip, while in experiment E_2 which is of the same tissue type as E_1, the D chip is brighter than the A chip. Thus, the challenge is that the data from the four separate chips in the Hu6800 format have to be normalized before data analysis on all the 7070 genes can be performed. Our goal is to combine the data from all the 16 experiments in the Hu6800 and the FL6800 chip formats.

On Affymetrix chips, each spot on the array is called a *probe*, which is a single stranded DNA oligonucleotide typically 20 - 30 bases long. Multiple probes are designed to hybridize to different regions of the *same* RNA. The set of probes that are used to detect one transcript is called the *probe set*. The HU chip format and the FL chip format consist of different probe sets, which is another key difference between the HU and FL chips. Hence, the signal intensities from the HU and FL chip formats are potentially different even after normalization.

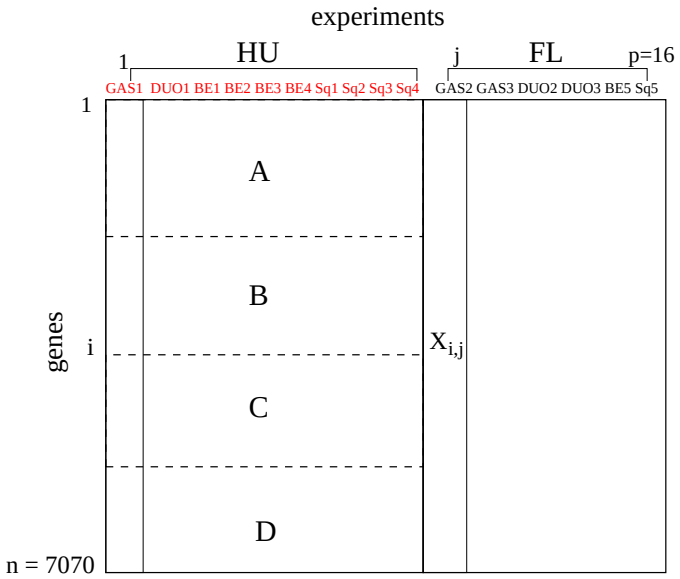

Figure 6.1: The Barrett's esophagus data set

6.2.1 Data normalization

The goal of this pre-processing step is to normalize the expression levels of the genes on separate A, B, C, D chips in the Affymetrix Hu6800 chips so as to perform data analysis on all the genes. Let $X_{i,j}$ denote the raw expression level (before normalization) of gene i under experiment j. To motivate the importance of this data normalization step, let us consider the following scenario. Let experiments E_1 and E_2 be experiments from the Hu6800 format. Suppose the overall expression levels of chip A in experiment E_1 are much higher than chip D in the same experiment E_1. Let gene g_i be a gene on chip A, and gene g_j be a gene on chip D, and suppose further that $X_{g_i,E_1} > X_{g_j,E_1}$. However, in experiment E_2, which is the same tissue type as experiment E_1, the overall expression levels of chip A is much lower than that of chip D under experiment E_2, and $X_{g_i,E_2} < X_{g_j,E_2}$. A discrepancy is observed: under the same tissue type, gene g_i is higher expressed than gene g_j under one experiment but not another. Without normalization, we cannot decide if the discrepancy is an artifact of chips A and D having different signal intensities under the two experiments E_1 and E_2, or a result of heterogeneity of tissue samples in the two experiments.

One difficulty of normalization is that the sets of genes on the four separate chips are mostly disjoint. There are only a few control genes that are in common among the four chips. We cannot obtain robust estimates of the mean and the standard deviation of chip intensity with only a few control genes.

The basic idea of our normalization approach is to use the data on the FL6800 format to determine the relative intensities of genes on each of the A,B, C, D chips in order to compare the expression levels of genes on different chips under the same experiment. The distributions of the raw expression levels $X_{i,j}$ from each experiment are highly skewed and have a very long tail. A typical example is shown in Figure 6.2, which is the histogram of the distribution of the expression levels in experiment Sq2. In the first step of normalization, we took the natural logarithm of all the expression levels from all 16 experiments. After the log transform, the distribution of the expression levels more closely resembles the normal distribution. The distribution of the log of the expression levels in experiment Sq2 is shown in Figure 6.3. Then, we normalized the log-transformed expression levels of each of the six experiments from the FL6800 format to mean 50 and standard deviation 10 (the choice of 50 and 10 is arbitrary). For each of the second set of experiments, E, from the FL6800 format (where E=BE5, Sq5, DUO2, DUO3, GAS2 or GAS3), the average expression levels μ_E^{chip} and standard deviations σ_E^{chip} (where $chip = A, B, C, D$) of corresponding genes on the A, B, C, D chips are computed. The expression levels of the first set of experiments were normalized to have the corresponding mean μ_E^{chip} and standard deviations σ_E^{chip} of the same tissue type. For example, the expression levels of genes in chip A from experiments Sq1, Sq2, Sq3 and Sq4 were scaled to have mean μ_{Sq5}^A and standard deviation σ_{Sq5}^A. The final distribution of experiment Sq2 is shown in Figure 6.4. In the case of duodenum, two experiments (DUO2, DUO3) were done on the FL6800 chips. The average of μ_{DUO2}^{chip} and μ_{DUO3}^{chip}, and the average of σ_{DUO2}^{chip} and σ_{DUO3}^{chip} (where $chip = A, B, C, D$) were used to normalize experiment DUO1. Similarly, GAS1 was normalized with the averages of GAS2 and GAS3.

After this normalization step, we can compare expression levels of genes across different chips from the first set of experiments. In terms of our motivating scenario, we can now compare the expression level of gene g_i on chip A to that of gene g_j on chip D. The normalized

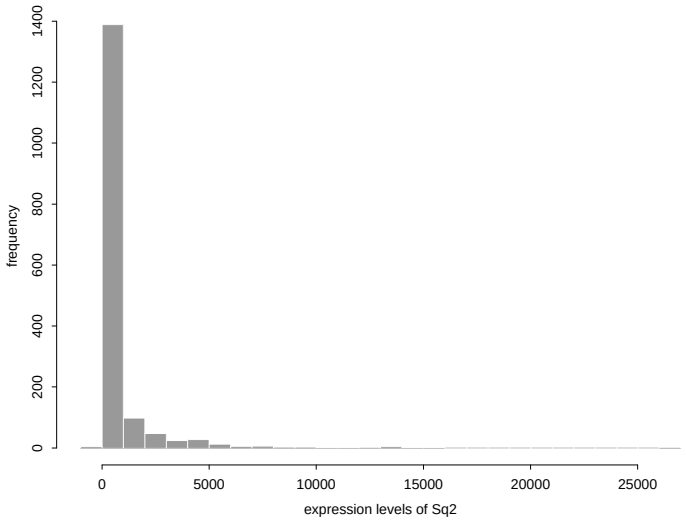

Figure 6.2: Histogram of the distribution of the expression levels in Sq2

data set is used in all subsequent analyses. The disadvantage of this approach is that even the same type of tissue samples can be heterogeneous, especially for neoplastic Barrett's epithelium.

6.2.2 Missing Data

Very low expression levels are usually marked with low confidence calls (the "absent" calls) by the Affymetrix software. In our analyses, we did not take into account any expression levels marked "absent" (low confidence) by the Affymetrix software. In the similarity analysis in Section 6.3, all the low confidence values are ignored. In computing the similarity of a pair of experiments, if the expression levels of gene g under one of the experiments is marked "absent", we ignore the expression levels from gene g. On the other hand, in the cluster analysis described in Section 6.4, expression values marked with the "absent" calls are thresholded with the minimum expression levels in the data set that are not "absent". Since at most 16 experiments contribute to the pairwise similarity of genes, ignoring the expression levels from some experiments would lead to unreliable pairwise similarity of genes. (This is in contrast to the thousands of genes that contribute to the pairwise similarity of experiments.)

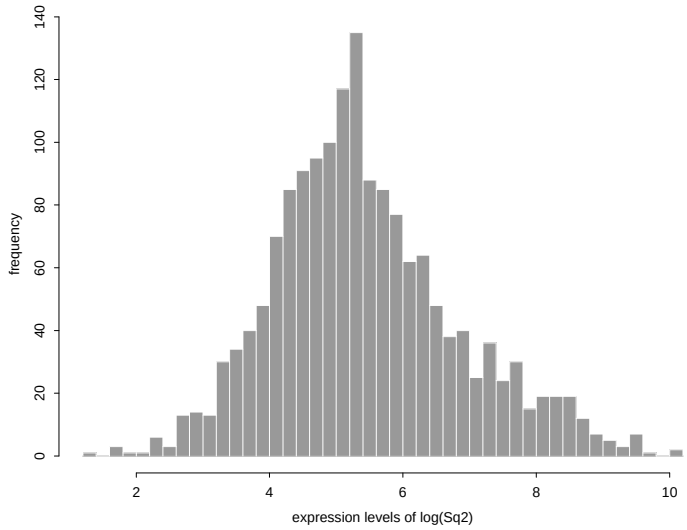

Figure 6.3: Histogram of the distribution of the expression levels after the natural log transformation in Sq2

6.3 Similarity between tissue samples

One of the goals of the Barrett's esophagus project is to investigate the distinction between neoplastic Barrett's epithelium and the surrounding normal tissues of the upper gastrointestinal tract at the expression level. We used Pearson's correlation coefficient [64] to quantify the pairwise similarities between tissue samples. Using the notations in Figure 6.1, the similarity (sample Pearson correlation coefficient) between experiment j and experiment k $(j, k = 1, \ldots, p)$ is

$$\frac{\sum_{g=1}^{n}(X_{g,j} - \mu_j)(X_{g,k} - \mu_k)}{\sqrt{\sum_{g=1}^{n}(X_{g,j} - \mu_j)^2 \sum_{g=1}^{n}(X_{g,k} - \mu_k)^2}} \tag{6.1}$$

where $\mu_j = \frac{\sum_{g=1}^{n} X_{g,j}}{n}$. The normalized data is used to compute the correlation coefficients.

Since we have multiple experiments on each tissue type, we averaged the normalized expression levels across experiments with the same tissue type in the same set of experiments in order to summarize the similarities between different tissue types. Then, Pearson's correlation coefficient was computed for each pair of tissue types and in each set of experiments. The results are shown in Table 6.1. Due to the different probe sets used by the Hu6800

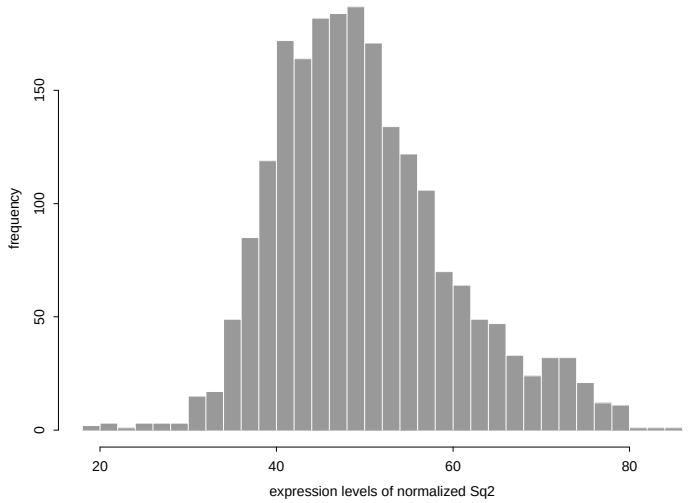

Figure 6.4: Histogram of the distribution of the expression levels after normalization in Sq2

and the FL6800 formats in the two sets of experiments, one must be careful in making comparisons between experiments from different chip formats. In Table 6.1, the notation E(r-s) (where $r < s$) means the average expression levels in experiments $Er, \ldots, Es$. For example, BE(1-4) represents the average expression levels in experiments BE1, BE2, BE3 and BE4.

The sample correlation coefficient is a *point estimate* of the true correlation coefficient between two tissue types, and hence it does not convey any uncertainty about the value of the estimate. Therefore, we also computed the 95% confidence intervals for the correlation coefficients to obtain a more robust comparison of the similarities between tissues. The confidence intervals can be computed using Fisher's "z transformation", which is a transformation of the correlation coefficient that was shown to be approximately normal [53]. Consequently, two non-overlapping confidence intervals suggest one pair of tissue types are more similar than the other pair with high probability.

The pairwise sample correlation coefficients from our first set of experiments (HU format) between the averaged normalized gastric and duodenum (0.807), gastric and squamous (0.751), and duodenum and squamous (0.732) showed that duodenum and gastric epithelium

Table 6.1: Average sample correlation coefficients between tissue types in the same set of experiments.

chip		*HU chips*				*FL chips*			
		GAS1	DUO1	BE(1-4)	Sq(1-4)	GAS(2-3)	DUO(2-3)	BE5	Sq5
HU	GAS1	1.000	0.807	0.851	0.751	0.864	0.763	0.805	0.741
HU	DUO1	0.807	1.000	0.841	0.732	0.761	0.872	0.792	0.719
HU	BE(1-4)	0.851	0.841	1.000	0.830	0.810	0.782	0.865	0.795
HU	Sq(1-4)	0.751	0.732	0.830	1.000	0.732	0.689	0.729	0.892
FL	GAS(2-3)	0.864	0.761	0.810	0.732	1.000	0.861	0.863	0.777
FL	DUO(2-3)	0.763	0.872	0.782	0.689	0.861	1.000	0.872	0.748
FL	BE5	0.805	0.792	0.865	0.729	0.863	0.872	1.000	0.796
FL	Sq5	0.741	0.719	0.795	0.892	0.777	0.748	0.796	1.000

are more related to each other at the expression level than either is to squamous. Furthermore, the confidence intervals for the correlation coefficients of gastric versus squamous epithelium ([0.730, 0.771]) and of duodenum versus squamous epithelium ([0.709, 0.753]) do not overlap with the confidence interval for gastric versus duodenum ([0.789, 0.824]), which further supports that duodenum and gastric are more similar to each other than either is to squamous. The results on our second set of experiments (FL format) are similar. The comparison of the expression profiles of the three normal gastrointestinal tissues is consistent with the more similar morphology and physiological role (secretory) of gastric and duodenal epithelia, when compared to the different morphology of non-secretory esophageal squamous epithelium.

For the first set of experiments, the sample correlation coefficients and confidence intervals between Barrett's epithelium and each of gastric epithelium (0.851, [0.839, 0.863]), duodenum (0.841, [0.827, 0.853]) and squamous epithelium (0.830, [0.817, 0.842]) showed that the sample correlation coefficients are comparable and the confidence intervals overlap. However, for the second set of experiments, Barrett's epithelium is more similar to gastric epithelium ($Sim(BE5, GAS(2-3)) = 0.863$), and duodenal epithelium ($Sim(BE5, DUO(2-3)) =$

0.872), than to squamous epithelium ($Sim(BE5, Sq5) = 0.796$). It turns out that this discrepancy is due to the fact that the averaged normalized expression levels are used. Table 6.2 shows the sample correlation coefficients between each of the individual 16 experiments without averaging the expression levels over the same tissue type. We also computed the 95% confidence interval (results not shown). From Table 6.2, experiment BE1 from the first set of experiments has lower similarity to gastric epithelium (GAS1) than to squamous epithelium (Sq1, Sq2, Sq3, Sq4). On the other hand, experiment BE4 (also from the first set of experiments) has higher similarity to gastric epithelium (GAS1) than to squamous epithelium (Sq1, Sq2, Sq3, Sq4). In the second set of experiments, experiment BE5 shows the same relative similarities as BE4, *i.e.,* $Sim(BE5, Sq5) < Sim(BE5, GAS2)$ and $Sim(BE5, Sq5) < Sim(BE5, GAS3)$. In fact, experiments BE4 and BE5 were derived from the same pooled tissue sample, but they were interrogated to different chip formats. Therefore, the discrepancy we observed using the average expression levels across tissue types in Table 6.1 merely reflects the heterogeneity of Barrett's epithelium. The overlapping confidence intervals for the correlation coefficients between individual pools of Barrett's epithelium with normal gastric, squamous and duodenum tissues (data not shown) suggest that Barrett's epithelium shared extensive transcriptional similarity with all of these surrounding normal tissues. Thus, there is no evidence for a Barrett's lineage-specific developmental association with one of the surrounding normal tissues. Several studies have shown that premalignant stages of BE contain different clonal populations of cells with multiple somatically acquired genetic abnormalities (for example, [7]). Therefore the variability in the expression patterns of Barrett's epithelium may reflect the genetic heterogeneity present in a neoplastic epithelium compared to surrounding normal tissues.

Figure 6.5 summarizes the sample correlation coefficients for the individual experiments from Table 6.2 by a dendrogram of hierarchical average-link clustering algorithm. Experiments under the same subtree in the dendrogram are more related to each other than experiments in a different subtree. Figure 6.5 shows that experiments of the same tissue type are more related to each other than to experiments of different tissue types despite the fact that different chip formats were used. In addition, experiment BE5 is most similar to experiment BE4 across all the experiments, even though they were interrogated to different

Table 6.2: Sample correlation coefficients between the individual experiments (not averaged over the same tissue types).

chip		*HU chips*										*FL chips*					
		GAS1	DUO1	BE1	BE2	BE3	BE4	Sq1	Sq2	Sq3	Sq4	GAS2	GAS3	DUO2	DUO3	BE5	Sq5
HU	GAS1	1.00	0.81	0.80	0.85	0.79	0.82	0.75	0.74	0.73	0.76	0.86	0.86	0.76	0.77	0.81	0.74
HU	DUO1	0.81	1.00	0.81	0.81	0.84	0.81	0.74	0.73	0.72	0.74	0.78	0.75	0.86	0.88	0.79	0.72
HU	BE1	0.80	0.81	1.00	0.89	0.83	0.84	0.81	0.81	0.80	0.82	0.77	0.75	0.73	0.75	0.79	0.76
HU	BE2	0.85	0.81	0.89	1.00	0.82	0.88	0.83	0.81	0.82	0.84	0.81	0.79	0.75	0.76	0.83	0.80
HU	BE3	0.79	0.84	0.83	0.82	1.00	0.81	0.74	0.73	0.71	0.76	0.79	0.76	0.76	0.79	0.84	0.72
HU	BE4	0.82	0.81	0.84	0.88	0.81	1.00	0.75	0.75	0.75	0.79	0.79	0.77	0.78	0.79	0.89	0.76
HU	Sq1	0.75	0.74	0.81	0.83	0.74	0.75	1.00	0.90	0.89	0.93	0.73	0.73	0.69	0.70	0.73	0.88
HU	Sq2	0.74	0.73	0.81	0.81	0.73	0.75	0.90	1.00	0.96	0.94	0.70	0.69	0.65	0.66	0.69	0.88
HU	Sq3	0.73	0.72	0.80	0.82	0.71	0.75	0.89	0.96	1.00	0.93	0.70	0.69	0.64	0.66	0.68	0.87
HU	Sq4	0.76	0.74	0.82	0.84	0.76	0.79	0.93	0.94	0.93	1.00	0.73	0.73	0.68	0.70	0.73	0.89
FL	GAS2	0.86	0.78	0.77	0.81	0.79	0.79	0.73	0.70	0.70	0.73	1.00	0.96	0.85	0.86	0.86	0.76
FL	GAS3	0.86	0.75	0.75	0.79	0.76	0.77	0.73	0.69	0.69	0.73	0.96	1.00	0.85	0.85	0.86	0.78
FL	DUO2	0.76	0.86	0.73	0.75	0.76	0.78	0.69	0.65	0.64	0.68	0.85	0.85	1.00	0.97	0.86	0.74
FL	DUO3	0.77	0.88	0.75	0.76	0.79	0.79	0.70	0.66	0.66	0.70	0.86	0.85	0.97	1.00	0.87	0.75
FL	BE5	0.81	0.79	0.79	0.83	0.84	0.89	0.73	0.69	0.68	0.73	0.86	0.86	0.86	0.87	1.00	0.80
FL	Sq5	0.74	0.72	0.76	0.80	0.72	0.76	0.88	0.88	0.87	0.89	0.76	0.78	0.74	0.75	0.80	1.00

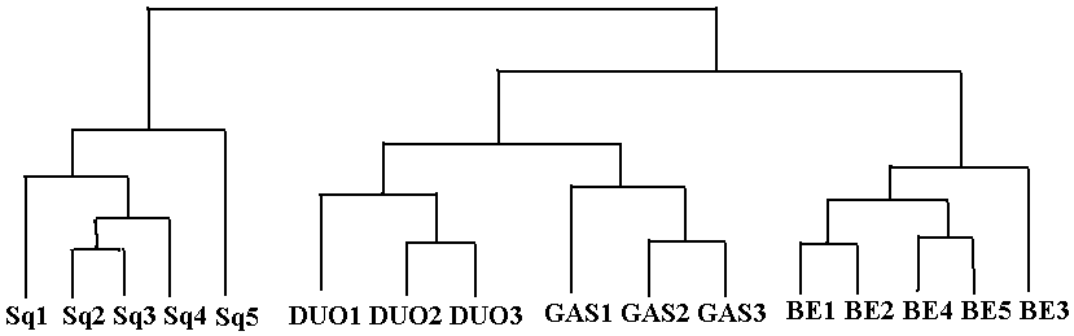

Figure 6.5: Dendrogram showing the relative similarities of the 16 experiments.

chip formats.

6.4 Cluster Analysis

In order to identify clusters of genes, the normalized data set is filtered to focus on genes that are differentially expressed in different tissue types. After we determined a set of differentially expressed genes, we applied the FOM methodology from Chapter 3 to choose a clustering algorithm. Finally, we applied the chosen clustering algorithm to produce clusters of genes with similar expression patterns.

6.4.1 Filtering

Our procedure to identify genes that are differentially expressed in different tissue types is similar to the standard procedure of the analysis of variance (ANOVA) [83]. Suppose we have independent samples from each of the k different populations, and the sample size from population i is n_i (where $i = 1, 2, \ldots, k$). Let $Y_{i,j}$ be an expression level from population i, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$. In the standard ANOVA procedure, $Y_{i,j}$'s are assumed to be independent, normal, $E[Y_{i,j}] = \mu_i$, $Var[Y_{i,j}] = \sigma^2$, and the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ versus $H_1 : H_0$ *is false* is tested. Note that the population variances are assumed to be equal. Let $n = n_. = \sum_i n_i$, $Y_{i.} = \sum_j Y_{ij}/n_i$, $Y_{..} = \sum_i Y_{i.} / \sum_i n_i = \sum_i \sum_j Y_{ij}/n$. The test statistic in the standard ANOVA procedure is the ratio of the between-population mean square to the residual mean square, *i.e.,*

$$\frac{\sum_i n_i (Y_{i.} - Y_{..})^2}{k-1} / \frac{\sum_i \sum_j (Y_{ij} - Y_{i.})^2}{n-k} \tag{6.2}$$

which follows the F-distribution with (k-1, n-k) degrees of freedom.

In the case of the Barrett's esophagus data, each tissue type is a population. There are four tissue types in our experiments: Barrett's epithelium, gastric epithelium, duodenal epithelium and squamous epithelium, *i.e.,* k is 4 and n is 16. The sizes of the tissue types, n_i, are 5, 5, 3, and 3 for Barrett's, squamous, gastric and duodenal epithelium respectively. For each gene, we tested the null hypothesis $H_0 : \mu_{BE} = \mu_{Sq} = \mu_{GAS} = \mu_{DUO}$ versus $H_1 : H_0$ *is false.* A gene is said to be *differentially expressed* if the null hypothesis H_0 is rejected.

Our idea is to use the test statistic in Equation 6.2, but instead of assuming that the test statistic follows the F-distribution with (k-1, n-k) degrees of freedom, an empirical distribution for the test statistic is computed. Due to the small sample sizes (3 or 5), the assumption of the F distribution can potentially have a large impact on the hypothesis testing. In the derivation of the test statistic, the normality assumption is used to show that the distribution of the test statistic in Equation 6.2 follows the F-distribution. Therefore, by generating an empirical distribution to compute the significance level, our approach does *not* assume the normality of the expression levels $Y_{i,j}$'s from each tissue type.

An empirical distribution for each gene is simulated by randomly permuting the expression levels of that gene from all the experiments, and by repeating the random permutation many times (3000 times in our implementation). If the test statistic of Equation 6.2 from the empirical distribution of a gene g is greater than the observed test statistic from the data less than 5% in all the random trials, we reject the null hypothesis H_0 at the 0.05 significance level. The gene g passes the filter, and is considered to be differentially expressed. Since the test statistic in Equation 6.2 is the ratio of the between tissue type mean square to the residual mean square, a large ratio implies that the population means are sufficiently different. Intuitively, our empirical testing procedure determines whether the observed ratio from the data is large enough so that it is not easily obtained by chance.

We applied the above modified ANOVA procedure to the thresholded normalized data

set. For 1095 genes (out of 7070 genes), the equal population mean null hypothesis is rejected at the 0.05 significance level, and hence pass the filter.

6.4.2 Choosing a clustering algorithm

With the filtered data set, the next problem is to choose a clustering algorithm for the data. We applied the figure of merit (FOM) methodology described in Chapter 3 to compare the performance of different clustering algorithms. The basic idea of the FOM methodology is to apply a clustering algorithm to the data from all but one experiment. The remaining experiment is used to assess the predictive power of the resulting clusters—meaningful clusters should exhibit less variation in the remaining experiment than clusters formed by chance. The predictive power of the resulting clusters is measured by the within-cluster variance, which is called the *figure of merit* (FOM). A clustering result with a small FOM implies low within-cluster variance, which in turn is an indication of high predictive power. The definition of FOM does not allow direct comparisons over different numbers of clusters. Therefore, the FOM is plotted against the number of clusters in typical FOM analyses.

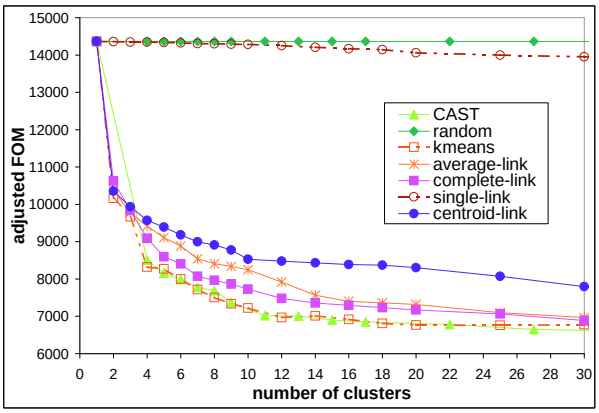

Figure 6.6: FOM analysis on the filtered Barrett's esophagus data (1095 genes).

Figure 6.6 shows the result of applying the FOM methodology to the filtered Barrett's esophagus data (1095 genes). As in Chapter 3, we used the correlation coefficient as the similarity metric. From Figure 6.6, single-link produces only slightly lower FOM than the random algorithm, which means that the performance of single-link is not satisfactory. The

k-means and CAST algorithms produce the lowest FOM, and have comparable performance. The FOM declines drastically up to around 8 clusters, so we estimate the number of clusters to be approximately 8.

With the FOM analysis in mind, we produced clustering results on the full data set (all 16 experiments) with 8 clusters using CAST, average-link and k-means. Then, we compared the clusters in light of prior biological knowledge about the data. Cytokeratins have a tissue specific profile and can be used to distinguish and purify tissues of interest in flow cytometry assays. Probe sets for twenty members of the cytokeratin genes passed the initial filtering criteria. The eight clusters generated by CAST and k-means correctly placed each of the cytokeratins in their respective clusters. In contrast, the average-link result, which produced a relatively higher FOM, did not correctly assign the cytokeratins to tissue specific clusters. Therefore, manual inspection and clusterings of cytokeratins suggest that CAST and k-means produce more robust clusters than average-link, which confirm our FOM analysis.

6.4.3 Clustering Results

We applied the CAST algorithm to the filtered Barrett's esophagus data set (1095 genes) to produce 8 clusters. Tissue specific clusters, in which the expression levels are relatively high in one tissue type, were obtained, for example, Figure 6.7 and Figure 6.8. (The other six clusters are not shown.) In Figures 6.7 and 6.8, the horizontal axis represents the 16 experiments, and the vertical axis shows the normalized expression levels. The solid line represents the average expression level in each experiment, and the dotted lines show one standard deviation above and below the average expression level in each cluster. Genes in the cluster in Figure 6.7 show relatively high expression levels in the five experiments using Barrett's epithelium tissue, while genes in the cluster in Figure 6.8 show relatively high expression levels in the five experiments using squamous epithelium tissue sample. Many interesting genes were found from these tissue specific clusters. The Barrett specific cluster (Figure 6.7) included genes associated with cell cycle progression (P1cdc47, PCM-1), cell migration (urokinase-type plasminogen receptor), growth regulation (TGF-beta super-

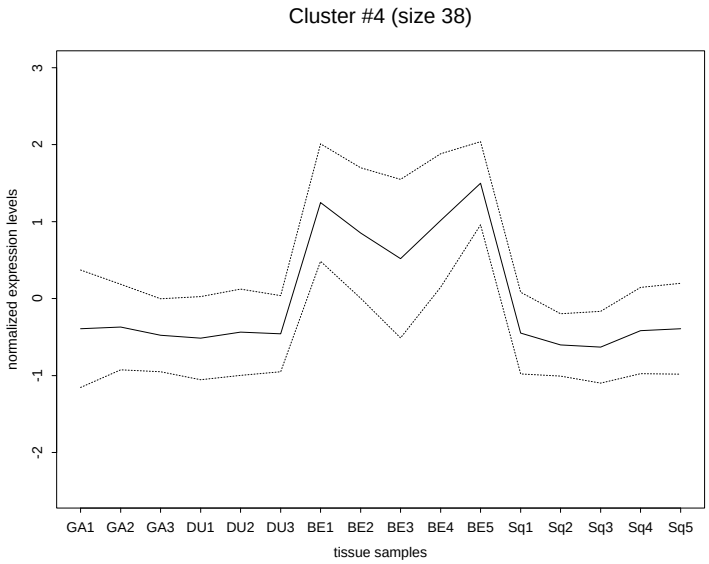

Figure 6.7: Barrett specific cluster

family, amphiregulin, Cyr61), stress responses (calcyclin, ATF3, TR3 orphan receptor) as well as epithelial cell surface antigens (epsilon-BP, Human surface antigen, integrin beta 4). The squamous specific cluster (Figure 6.8) included oncogenes (pim-1, met, P47 LBC), a number of proteinase inhibitors (maspin, elafin, monocyte/neutrophil elastase inhibitor, cystatin M, cystatin B, squamous cell carcinoma antigen, urokinase inhibitor), proteases (protease M, calcium dependent protease) and a series of small proline rich proteins (sprI, sprII, SPRR2B, SPR2-1, SPRR1A) implicated in various cellular stress responses. For more detailed biological interpretation, please refer to our paper [8].

A careful inspection of the clusters in Figures 6.7 and 6.8 shows that the experiments using the same pool of tissue samples (BE4 and BE5, Sq2 and Sq5) do not have identical normalized expression levels. The differences between the normalized expression levels of the same tissue samples hybridized to both HU6800 and FL6800 chips reflect the experimental variation and signal intensity variation from different chip formats.

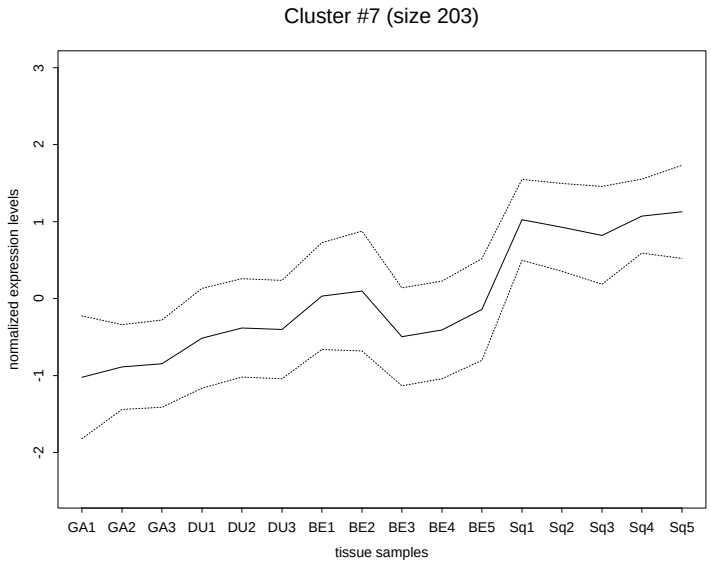

Figure 6.8: Squamous specific cluster

6.5 Summary and Future Work

We demonstrated the application of data normalization, correlation analysis, data filtering, the FOM methodology and clustering algorithm to the Barrett's esophagus data. In many papers on gene expression analysis, the issues of normalization and data pre-processing are not discussed in detail. In particular, we are not aware of any work addressing the issue of different average intensities on separate chips (A, B, C, D chips) from the same experiment. We feel that normalization and data pre-processing should not be overlooked, and should deserve as much attention as any subsequent analysis. In our cluster analysis, we replaced all expression values with "absent" calls with a threshold value. This is a preliminary way to handle missing data values. A possible direction of future work would be to investigate the effect of different missing data handling methods on subsequent analyses on the data.

We applied many different heuristic-based clustering algorithms on the Barrett's esophagus data. It would also be interesting to study the clustering results using the model-based approach described in Chapter 5. However, the missing data problem needs to be solved before we can apply the model-based approach because the current method of thresholding the "absent" data values would lead to violations of the Gaussian mixture assumptions.

There is a large statistics literature on incomplete data (for example, Schafer [70]), so one direction of future work would be to study how the Gaussian mixture assumption is satisfied or violated using different statistical methods on incomplete data, and to test the methods on the Barrett's esophagus data.

In our filtering procedure, the expression levels of the experiments corresponding to each tissue type were assumed to be independent. This is not the case for the experiments using the same pool of tissue samples (*i.e.*, BE4 and BE5, Sq2 and Sq5). It would be interesting to modify our current approach to account for the dependence between tissue samples.

BIBLIOGRAPHY

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

[2] J. Aitchison. *The Statistical Analysis of Compositional Data.* Chapman and Hall, 1986.

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

[4] D. F. Andrews, R. Gnanadesikan, and J. L. Warner. Methods for assessing multivariate normality. In P. R. Krishnaiah, editor, *Multivariate analysis III*, pages 95–116. New York: Academic Press, 1973.

[5] J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[6] Y. Barash and N. Friedman. Context-specific Bayesian clustering for gene expression data. In T. Lengauer et al., editors, *RECOMB 2001, Proceedings of the Fifth Annual International Conference on Computational Biology*, pages 12–21, Montreal, Canada, April 2001.

[7] M. T. Barrett, C. A. Sanchez, L. J. Prevo, D. J. Wong, P. C. Galipeau, T. G. Paulson, P.S Rabinovitch, and B. J. Reid. Evolution of neoplastic cell lineages in Barrett oesophagus. *Nature Genetics*, 22(1):106–109, 1999.

[8] M. T. Barrett, K. Y. Yeung, W. L. Ruzzo, L. Hsu, P. L. Blount, R. Sullivan, H. Zarbl,

J. Delrow, P. S. Rabinovitch, and B. J. Reid. Transcriptional analysis of Barretts metaplasia and normal upper GI mucosae. To appear in Neoplasia, 2001.

[9] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6:281–297, 1999.

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

[11] W. J. Blot, S. S. Devesa, R. W. Kneller, and J. F. Jr. Fraumeni. Rising incidence of adenocarcinoma of the esophagus and gastric cardia. *Journal of the American Medical Association*, 265(10):1287–1289, 1991.

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

[13] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[14] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

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

[16] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *The Journal of the Pattern Recognition Society*, 28:781–793, 1995.

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

[18] T. Chen, V. Filkov, and S. S. Skiena. Identifying gene regulatory networks from experimental data. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 94–103, Lyon, France, March 1999.

[19] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proc. Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, La Jolla, USA, August 2000.

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

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

[22] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

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

[24] W. H. E. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1:7–24, 1984.

[25] D. Defays. An efficient algorithm for a complete link method. *Computer Journal*, 20:364–366, 1977.

[26] B. Efron. *The jackknife, the bootstrap, and other resampling plans.* Society for Industrial and Applied Mathematics, 1982.

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

[28] B. S. Everitt. *Clustering Analysis.* John Wiley and Sons, 1993.

[29] C. Fraley. Algorithms for Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270–281, 1998.

[30] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

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

[32] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[33] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.

[34] S. R. Hamilton, R. R. Smith, and J. L. Cameron. Prevalence and characteristics of Barrett esophagus in patients with adenocarcinoma of the esophagus or esophagogastric junction. *Human Pathology*, 19(8):942–948, 1988.

[35] J. A. Hartigan. *Clustering Algorithms.* John Wiley and Sons, 1975.

[36] Robert V. Hogg and Allen T. Craig. *Introduction to Mathematical Statistics.* Macmillan Publishing Co., Inc., New York, NY, 4 edition, 1978.

[37] I. Holmes and W. J. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In R. Altman et al., editors, *Proceedings Eighth Annual International Conference on Intelligent Systems for Molecular Biology*, pages 202–210, La Jolla, CA, August 2000. AAAI Press.

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

[39] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

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

[41] A. K. Jain and J. V. Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5):547–568, 1987.

[42] J. D. Jobson. *Applied multivariate data analysis.* New York: Springer-Verlag, 1991.

[43] I. T. Jolliffe. *Principal Component Analysis.* New York : Springer-Verlag, 1986.

[44] I. T. Jolliffe, B. Jones, and B. J. T. Morgan. Cluster analysis of the elderly at home: a case study. *Data analysis and Informatics*, pages 745–757, 1980.

[45] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[46] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, New York, 1990.

[47] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1143, 1995.

[48] T. Kohonen. *Self-organizing maps.* Springer-Verlag, 1997.

[49] E. S. Lander. Array of hope. *Nature Genetics*, 21:3–4, 1999.

[50] L. Lazzeroni and A. B. Owen. Plaid models for gene expression data. Technical Report 211, Dept. of Statistics, Stanford University, 2000.

[51] D. S. Levine, B. J. Reid, R. C. Haggitt, C. E. Rubin, and P. S. Rabinovitch. Correlation of ultrastructural aberrations with dysplasia and flow cytometric abnormalities in Barrett's epithelium. *Gastroenterology*, 96:355–367, 1989.

[52] D. S. Levine, C. E. Rubin, B. J. Reid, and R. C. Haggitt. Specialized metaplastic columnar epithelium in Barrett's esophagus. A comparative transmission electron microscopic study. *Laboratory Investigation*, 60(3):418–432, 1989.

[53] A. M. Liebetrau. *Measures of Association.* Sage Publications, 1983.

[54] R. J. Lipshutz, S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21:20–24, 1999.

[55] J. MacQueen. Some methods for classication and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1965.

[56] K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970.

[57] L. A. Mather. A linear algebra measure of cluster quality. *Journal of the American Society for Information Science*, 51:602–613, 2000.

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

[59] G. J. McLachlan and D. Peel. *Finite Mixture Models*. New York: Wiley, 2000.

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

[61] G. S. Michaels, D. B. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi. Cluster analysis and data visualization of large-scale gene expression data. In *Pacific Symposium on Biocomputing 3*, pages 42–53, 1998.

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

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

[64] K. Pearson. Mathematical contributions to the theory of evolution, III. Regression, heredity, and panmixia. *Philosophical Transcriptions of the Royal Society*, A 187:253–318, 1896.

[65] R. W. Phillips and R. K. Wong. Barrett's esophagus. Natural history, incidence, etiology, and complications. *Gastroenterology Clinics of North America*, 20(4):791–816, 1991.

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

[67] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, volume 5, pages 452–463, 2000.

[68] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[69] J. A. Salo, E. O. Kivilaakso, T. A. Kiviluoto, and I. O. Virtanen. Cytokeratin profile suggests metaplastic epithelial transformation in Barrett's oesophagus. *Annals of Medicine*, 28(4):305–309, 1996.

[70] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, 1997.

[71] M. Schummer. Manuscript in preparation. 2000.

[72] M. Schummer, W. V. Ng, R. E. Bumgarner, P. S. Nelson, B. Schummer, D. W. Bednarski, L. Hassell, R. L. Baldwin, B. Y. Karlan, and L. Hood. Comparative hybridization of an array of 21500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Genes*, 238:375–385, 1999.

[73] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[74] S. Z. Selim and M. A. Ismail. K-means type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1):81–86, 1984.

[75] R. Shamir and R. Sharan. Algorithmic approaches to clustering gene expression data. In *Current Topics in Computational Biology*. MIT Press, 2001.

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

[77] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. 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, 1999.

[78] S. Tavazoie, J. D. Huges, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

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

[80] L. Toldo. Cluster validity. Personal communications and poster presentation at the Third International Meeting on Microarray Data Standards, Annotations, Ontologies and Databases, 2001.

[81] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Science USA*, 95:334–339, January 1998.

[82] R. A. Young. Biomedical discovery with DNA arrays. *Cell*, 102(1):9–15, 2000.

[83] J. H. Zar. *Biostatistical Analysis.* Prentice Hall, 1984.

[84] L. P. Zhao, R. Prentice, and L. Breeden. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proceedings of the National Academy of Science USA*, 98:5631–5636, 2001.

[85] G. K. Zipf. *Human behavior and the principle of least effort.* Addison-Wesley, 1949.

Appendix A

SELF-ORGANIZING MAPS

The *self-organizing map* (SOM) is a popular clustering algorithm for gene expression data [77]. In Chapter 3, we compared the performance of a few popular clustering algorithms on various gene expression data. Many of our colleagues and collaborators suggested that it would be interesting to compare the performance of SOM as well. We decided not to consider SOM in our comparisons because there are many parameters to be tuned in SOM. Our preliminary investigation of the stability of SOM with respect to different parameters showed that clustering results from SOM are not consistent when the parameters are varied.

The SOM defines a mapping from the input data space $\Re^p$ onto a two dimensional array of nodes [48]. Every node i is associated with a *reference vector* $m_i \in \Re^p$. The *topology* of the array of nodes is usually either rectangular or hexagonal, which is one of the parameters that need to be specified by the users. In addition, the users also have to specify the size and geometry of the topology. In the case of a rectangular array of nodes, the size should be stated in terms of the number of nodes horizontally and vertically. Figure A.1 is an example of a 3 by 2 rectangular grid. One major difficulty of applying our FOM methodology in Chapter 3 using SOM is that there are many possible geometries for a given number of clusters even with the topology fixed. For example, twenty clusters can be represented in a 1 by 20 or 2 by 10 or 4 by 5 rectangular grid.

The SOM can be considered as a projection of the probability density function of the high dimensional input data onto this two dimensional array of nodes. An input vector $x \in \Re^p$ is compared with each of the reference vectors m_i, and the input vector is mapped to the node with the reference vector that best matches the input vector. The SOM is usually initialized with random reference vectors. After initialization, the map is trained through a learning process: nodes that are topographically close in the array up to a certain distance will activate each other to learn from the same input vector. In the learning process, the

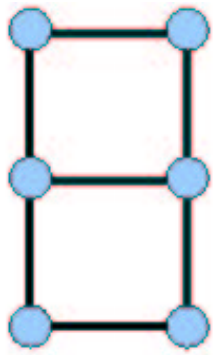

Figure A.1: An example of a 3 by 2 rectangular grid.

learning rate and the distance in which nodes activate each other also need to be specified by the users.

In our experiments, we used the yeast cell cycle data described in Chapter 2, and the SOM package from Kohonen [48]. We measured the consistency of clustering results using the average adjusted Rand index (described in Chapter 2) over pairs of clustering results from different parameter settings in ten random initializations. Using the hexagonal topology and the same distance and learning parameters, comparing clustering results of grid size 4 by 5 to 1 by 20 yields an average adjusted Rand index of 0.453, and comparing clustering results of grid size 4 by 5 to 2 by 10 yields an average adjusted Rand index of 0.451. In some software implementations (for example, GeneCluster by Tamayo *et al.* [77]), default parameters for the distance and learning parameters are given. Our preliminary experimental results showed that even if the topology, number of clusters, the distance and learning parameters are fixed, it is still not trivial to compare clustering results from SOM with different geometry.

Appendix B

CORRELATION COEFFICIENT WHEN THERE ARE 2 COMPONENTS

When there are 2 components, the correlation coefficient is either 1 or -1. Suppose there are two genes g_1 and g_2 with two components. Let $x_{i,j}$ (where $i, j = 1, 2$) be the expression level of gene i under component j. The correlation coefficient between g_1 and g_2 can be simplified to:

$$\frac{(x_{1,1} - x_{1,2}) * (x_{2,1} - x_{2,2})}{\sqrt{(x_{1,1} - x_{1,2})^2 * (x_{2,1} - x_{2,2})^2}} \tag{B.1}$$

Since the denominator in Equation B.1 represents the product of the norms of genes g_1 and g_2, the denominator must be positive. From Equation B.1, the correlation coefficient between genes g_1 and g_2 is 1 if $(x_{1,1} - x_{1,2}) * (x_{2,1} - x_{2,2}) > 0$, the correlation coefficient is -1 if $(x_{1,1} - x_{1,2}) * (x_{2,1} - x_{2,2}) < 0$. If $x_{1,1} = x_{1,2}$ or $x_{2,1} = x_{2,2}$, the correlation coefficient is undefined. Since there are only two possible values that the correlation coefficient can take when there are two components, there are at most two clusters.

VITA

Ka Yee Yeung studied computer science at University of Waterloo in Ontario, Canada (B.Math 1995 and M.Math 1996). She moved to Seattle in 1996 to continue her graduate work in computer science at University of Washington (M.S. 1998, Ph.D. 2001).

During her graduate career at University of Washington, she became interested in computational biology, especially gene expression analysis. As a graduate student at University of Washington, she was supervised by Professor Richard M. Karp and Professor Walter L. Ruzzo.