

# Use of SVD-based probit transformation in clustering gene expression profiles

Faming Liang\*

*Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA*

Received 19 February 2006; received in revised form 20 January 2007; accepted 23 January 2007  
Available online 16 February 2007

---

## Abstract

The mixture-Gaussian model-based clustering method has received much attention in clustering gene expression profiles in the literature of bioinformatics. However, this method suffers from two difficulties in applications. The first one is on the parameter estimation, which becomes difficult when the dimension of the data is high or the size of a cluster is small. The second one is on the normality assumption for gene expression levels, which is seldom satisfied by real data. In this paper, we propose to overcome these two difficulties by the probit transformation in conjunction with the singular value decomposition (SVD). SVD reduces the dimensionality of the data, and the probit transformation converts the scaled eigensamples, which can be interpreted as correlation coefficients as explained in the text, into Gaussian random variables. Our numerical results show that the SVD-based probit transformation enhances the ability of the mixture-Gaussian model-based clustering method for identifying prominent patterns of the data. As a by-product, we show that the SVD-based probit transformation also improves the performance of the model-free clustering methods, such as hierarchical,  $K$ -means and self-organizing maps (SOM), for the data sets containing scattered genes. In this paper, we also propose a run test-based rule for selection of eigensamples used for clustering.

© 2007 Elsevier B.V. All rights reserved.

*Keywords:* Gene expression profiles; Model-based clustering; Probit transformation; Singular value decomposition

---

## 1. Introduction

In attempts to understand biological systems, large amount of gene expression data have been generated by researchers. Due to the large number of genes and the complexity of the biological systems, clustering has been one of the most important exploratory tools for analyzing these data. Clustering identifies groups of genes that exhibit similar expression profiles. The popular clustering methods can be divided into two categories, namely, model-free methods and model-based methods. In the model-free clustering methods, no probabilistic model is specified for the data, and clustering is made either by optimizing a certain target function or iteratively agglomerating/dividing genes to form a bottom-up/top-down tree. Examples include  $K$ -means (Tavazoie et al., 1999; Tou and Gonzalez, 1979), hierarchical clustering (Carr et al., 1997; Eisen et al., 1998), self-organizing maps (Tamayo et al., 1999), among others. The model-based clustering methods construct clusters based on the assumption that the data follows a mixture distribution. A non-exhaustive list of recent works in this direction include Banfield and Raftery (1993), Biernacki et al. (1999), Fraley and Raftery (2002), Yeung et al. (2001), Medvedovic and Sivaganesan (2002), McLachlan et al. (2002),

---

\* Tel.: +1 979 8453197; fax: +1 979 8453144.  
E-mail address: [fliang@stat.tamu.edu](mailto:fliang@stat.tamu.edu).

Wakefield et al. (2003) and Medvedovic et al. (2004). One advantage of the model-based clustering methods is that the probability model can be used in criteria to choose an appropriate number of clusters.

Among the model-based clustering methods, the one based on the mixture-Gaussian distribution is much more interesting due to its simplicity in computation. Henceforth, the mixture-Gaussian model-based clustering method will be abbreviated as the MG method. The MG method assumes that the observations  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are generated from a mixture Gaussian distribution with an unknown number of components. The corresponding likelihood function is

$$L(\mathbf{x}_1, \dots, \mathbf{x}_n | \omega_k, \mu_k, \Sigma_k, k = 1, \dots, G) = \prod_{i=1}^n \left[ \sum_{k=1}^G \frac{\omega_k}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-1/2(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)} \right], \quad (1)$$

where  $G$  is the (unknown) number of components,  $p$  is the dimension of the observations,  $\omega_k$  is the probability that an observation belongs to component  $k$  ( $\omega_k \geq 0$  and  $\sum_{k=1}^m \omega_k = 1$ ), and  $\mu_k$  and  $\Sigma_k$  are the mean vector and covariance matrix of component  $k$ , respectively. Banfield and Raftery (1993) proposed to reparametrize the covariance matrices by eigenvalue decomposition in the form:

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \quad (2)$$

where  $\lambda_k = |\Sigma_k|^{1/d}$ ,  $D_k$  is the matrix of eigenvectors of  $\Sigma_k$ , and  $A_k$  is a diagonal matrix such that  $|A_k| = 1$ . The parameter  $\lambda_k$  determines the volume of component  $k$ ,  $D_k$  determines its orientation and  $A_k$  its shape. Allowing some but not all of these quantities to vary between components results in a set of parsimonious models which are appropriate to describe various clustering situations. Fraley and Raftery (2002) considered 10 different models related to different assumptions on the component variance matrices. Each model is denoted by a string of three letters E (equal), I (identical) and V (variable). The first letter of the string states the assumption on the volumes of the clusters, the second letter on the shapes, and the third letter on the orientation. For example, VEI represents a model in which the volumes of clusters may vary (V), the shapes of all clusters are equal (E), and the orientation is the identity (I). The MG method is implemented in the software MCLUST, which is downloadable at <http://www.stat.washington.edu/mclust>. In MCLUST, the model parameters are estimated using the EM algorithm (Dempster et al., 1977), and the BIC criterion is adopted for determining the number of clusters and the covariance structure.

Although the MG method has achieved great successes in clustering gene expression profiles (Yeung et al., 2001), its applications may be seriously hindered by the following two difficulties. The first one is on the parameter estimation, which becomes difficult when the dimension of the data is high or the size of a cluster is small. The second one is on the validity of the normality assumption. This assumption is seldom satisfied by real data. Applying the MG method to a data set with distribution deviated from Gaussian will result in a sub-optimal clustering result.

In this paper, we propose to overcome the above two difficulties by a SVD-based probit transformation. SVD reduces the dimension of the observations, and the probit transformation converts the scaled eigensamples, which can be interpreted as correlation coefficients as explained below, into Gaussian random variables. Our numerical results show that the transformation enhances the ability of the MG method for identifying prominent patterns of the data. In this paper, we also propose a run test-based rule for selection of eigensamples used for clustering. The new rule works well for all examples studied in this paper. Although the main theme of this paper is to show that the SVD-based probit transformation generally improves the performance of the MG method in clustering gene expression profiles, as a by-product we show that the transformation also improves the performance of the model-free clustering methods for the data sets containing scattered genes (Tseng and Wong, 2005).

The remaining part of this paper is organized as follows. In Section 2, we describe the SVD-based probit transformation and illustrate it using two motivating examples. In Section 3, we test the performance of the transformation on three simulated examples. In Section 4, we apply the transformation to a real data example. In Section 5, we conclude the paper with a brief discussion.

## 2. SVD-based probit transformation for gene expression profiles

### 2.1. SVD-based probit transformation

Principal component analysis, or its computational equivalent the SVD analysis, has long been considered as a useful tool for reducing the dimensionality of the data prior to clustering, see, for example, Jolliffe (1986). Recently, this tool

has been applied to gene expression data (Holter et al., 2000; Alter et al., 2000; Yeung and Ruzzo, 2001; Hastie et al., 2000; Horn and Axel, 2003).

Let  $X$  denote an  $n \times p$  matrix, which represents a data set of  $n$  genes with each being measured at  $p$  discrete time points. The  $i$ th row of  $X$ , denoted by  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ , is called the expression profile of gene  $i$ . Further, we suppose that each profile has been normalized to have unit length, i.e.,  $\|\mathbf{x}_i\| = \sqrt{\sum_j x_{ij}^2} = 1$ . The normalization suppresses the magnitude and enhances the pattern of the gene expression profiles. For some examples, we also follow the literature to normalize  $\mathbf{x}_i$ 's to have mean 0 and unit length. If this normalization is used, it will be stated explicitly in the text. According to the SVD theorem (Watkins, 1991), the matrix  $X$  can be decomposed in the form

$$X = U\Lambda V^T, \tag{3}$$

where  $U$  and  $V$  are  $n \times \min(n, p)$  and  $p \times \min(n, p)$  matrices of orthonormal columns, and  $\Lambda$  is a diagonal matrix of ordered non-negative singular values. Let  $\sigma_1 \geq \dots \geq \sigma_r > 0$  denote the  $r$  non-zero diagonal elements of  $\Lambda$ , where  $r$  is the rank of the matrix  $X$ , and  $\sigma_i$ s are equal to the square roots of  $r$  non-zero eigenvalues of  $XX^T$  and  $X^T X$ . The columns of  $V$ , denoted by  $\mathbf{v}_1, \dots, \mathbf{v}_r$ , are called eigenpatterns. The columns of  $U$ , determined by the formula  $\mathbf{u}_i = (1/\sigma_i)X\mathbf{v}_i$  for  $i = 1, \dots, r$ , are called eigensamples. The rows of  $U$  are called eigengenes. To reduce the dimensionality of the data, a subset of eigensamples are often used in clustering in place of  $X$ . The underlying rationale is that the eigensamples have extracted the clustering structure of the data. A variety of rules, some of them are informal and ad hoc, have been proposed for determining the subset of eigensamples that can be used for clustering. Refer to Jolliffe (1986) for an overview of these rules. In this paper, a non-parametric test-based rule is proposed for selection of eigensamples.

Although the matrix  $U$  can be used in a variety of model-free clustering methods, such as  $K$ -means and hierarchical as demonstrated by Yeung and Ruzzo (2001), it cannot be used directly in the MG method. This is because the matrix  $U$  is actually a scaled correlation coefficient matrix. The  $(i, j)$  entry of the matrix  $XV (=UA)$  can be interpreted as the (non-central) Pearson coefficient of the  $i$ th gene expression profile and the  $j$ th eigenpattern. Recall that each row of  $X$  has been normalized to have unit length and  $V$  is a matrix of orthonormal columns. To cluster eigengenes using the MG method, we propose to use the probit transformation to convert them to Gaussian random variables. Let  $U^* = (u_{ij}^*) = XV$ . The probit transformation converts  $u_{ij}^*$  to  $z_{ij}$ ,

$$z_{ij} = \Phi^{-1}((1 + u_{ij}^*)/2), \quad i = 1, \dots, n, \quad j = 1, \dots, r, \tag{4}$$

where  $\Phi$  denotes the CDF of the standard Gaussian distribution. Each row of  $Z = (z_{ij})$  is called a transformed eigengene, and each column a transformed eigensample.

Fig. 1 depicts the probit transformation. It tends to convert a correlation coefficient near  $-1$  toward  $-\infty$ , and convert a correlation coefficient near  $1$  toward  $+\infty$ . While for a correlation coefficient near  $0$ , it tends to convert it linearly. Different curvatures of the transformation at different values of the correlation coefficient make the transformation potentially useful in separating different clusters of the genes. This point will be further explored in the next subsection.

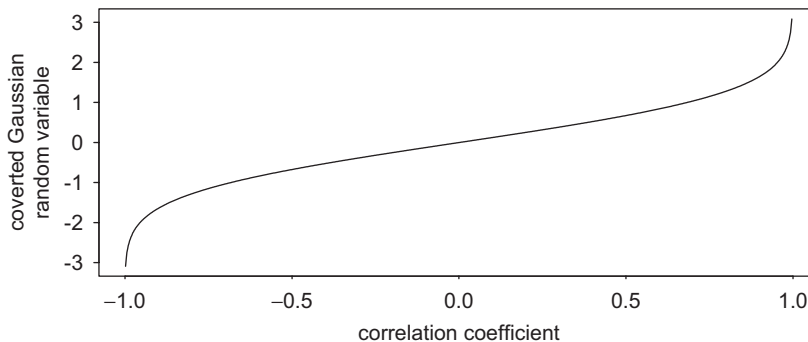


Fig. 1. A depiction of the probit transformation.

## 2.2. An illustrative example

This example consists of 10 data sets. Each consists of 1000 genes, among which 400 genes are generated from a six-dimensional Gaussian distribution  $N_6(\boldsymbol{\mu}, 0.3^2 I_6)$ , 300 genes from  $N_6(-\boldsymbol{\mu}, 0.3^2 I_6)$ , and 300 genes from  $N(\mathbf{0}, 0.3^2 I_6)$ , where  $\boldsymbol{\mu} = (\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, -\frac{3}{4}, -\frac{3}{4}, -\frac{3}{4})$  and  $I_6$  is the six-dimensional identity matrix. Each gene is normalized to have mean 0 and unit length. This example mimics the avian pineal gland expression data analyzed by Liang and Wang (2007).

In each data set, the genes generated from the distribution  $N(\mathbf{0}, 0.3^2 I_6)$  can be regarded as scattered genes. In microarray experiments, the scattered genes refer to the genes whose expression levels do not change in response to the experimental conditions. Reflecting in observations, the expression levels of the scattered genes do not change much across samples and often show low correlation to the expression levels of any unscattered genes. If the scattered genes are forced into a cluster, the average profile of this cluster can be compromised. Numerical results show that many of the popular clustering methods, such as hierarchical,  $K$ -means and self-organizing maps, tend to fail to separate the scattered and unscattered genes, and the patterns identified by them tend to be contaminated by scattered genes. To avoid this problem, Tamayo et al. (1999) suggested to filter out scattered genes prior to clustering. For example, they set a filter for the yeast expression data (Cho et al., 1998) to eliminate the genes which did not show a relative change of 2 and an absolute change of 35 units. However, due to the complexity of the gene expression mechanism, it is impossible to set a filter which can avoid simultaneously the two types of errors, removing some non-scattered genes from the data set and leaving some scattered genes in the data set. Hence, a clustering method robust to the existence of scattered genes is of interest to us.

Now, we show that the SVD-based probit transformation can generally improve the performance of many clustering methods for the data sets containing scattered genes. Fig. 2 illustrates the probit transformation for one data set generated in this example. Fig. 2(a) shows the first eigenpattern of the data. It captures the main expression pattern of the genes. Fig. 2(b) shows the histogram of the transformed first eigensample. It suggests that the genes in the data set can be grouped into three clusters, which correspond to the right, left and middle modes of the histogram, respectively. The respective patterns are parallel, anti-parallel and uncorrelated with the first eigenpattern. With only the transformed first eigensample, the MG method has already been able to separate the three clusters approximately. Fig. 2(c) is the scatter plot of the genes in the subspace of the first eigensample and the second eigensample. Fig. 2(d) is the scatter plot of the genes in the subspace of the transformed first eigensample and the transformed second eigensample. It shows that the probit transformation squeezes (relatively) the cluster 3 toward the center (0,0), while pulls out clusters

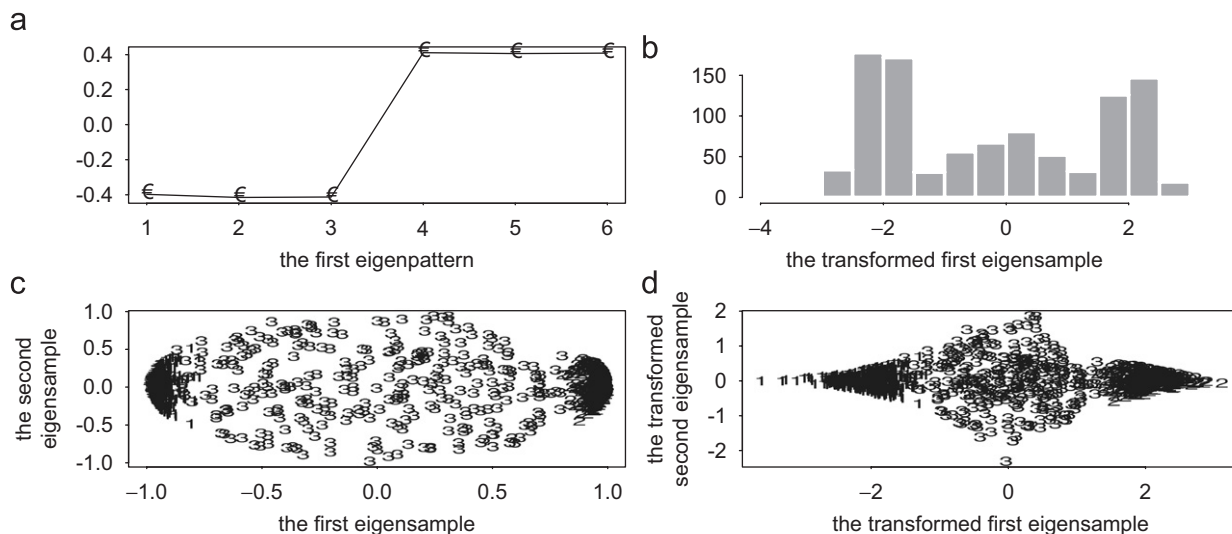


Fig. 2. Probit transformation for one data set. (a) The first eigenpattern. (b) The histogram of the transformed first eigensample. (c) The scatter plot of the genes in the subspace of the first eigensample and the second eigensample. (d) The scatter plot of the genes in the subspace of the transformed first eigensample and the transformed second eigensample. The numbers 1, 2 and 3 in plots (c) and (d) indicate the clusters of the genes.

1 and 2 toward  $-\infty$  and  $\infty$ , respectively. The probit transformation benefits many clustering methods, especially for those distance-based methods, such as the hierarchical and  $K$ -means methods.

In this paper, we follow [Yeung and Ruzzo \(2001\)](#) and [Yeung et al. \(2001\)](#) to use the adjusted Rand index ([Rand, 1971](#); [Hubert and Arabie, 1985](#)) to assess the quality of a clustering result when the true clusters are known. The adjusted Rand index measures the degree of agreement between two partitions of the same set of observations, even when the comparing partitions have different numbers of clusters. Let  $\Omega$  denote a set of  $n$  observations,  $C = \{c_1, \dots, c_s\}$  and  $C' = \{c'_1, \dots, c'_t\}$  represent two partitions of  $\Omega$ ,  $n_{ij}$  be the number of observations that are in both cluster  $c_i$  and cluster  $c'_j$ ,  $n_i$  be the number of observations in cluster  $c_i$ , and  $n_j$  be the number of observations in cluster  $c'_j$ . The adjusted Rand index can be calculated as

$$\rho = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[ \sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2} \right] / \binom{n}{2}}{\left[ \sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2} \right] / 2 - \left[ \sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2} \right] / \binom{n}{2}}. \tag{5}$$

A higher value of  $\rho$  means a higher correspondence between the two partitions. When the two partitions are identical,  $\rho$  is 1. When a partition is random, the expectation of  $\rho$  is 0. Under the generalized hypergeometric model, it can be shown ([Hubert and Arabie, 1985](#)) that

$$E \left[ \sum_{i,j} \binom{n_{ij}}{2} \right] = \left[ \sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2} \right] / \binom{n}{2}. \tag{6}$$

Refer to [Hubert and Arabie \(1985\)](#) for the theoretical properties of  $\rho$ .

The MG method was first applied to this example. To illustrate the effect of the probit transformation, we did not reduce the dimensionality of the data for this example; that is, we used all transformed eigensamples in clustering the genes. The BIC analysis suggests the VVI model for the transformed data. We here assume that the true number of clusters is known. This assumption is also applied to other methods considered below. [Fig. 3](#) shows the true clusters (plots (a)–(c)) and the clusters (plots (d)–(f)) obtained by the MG method for one data set. Totally there are 23

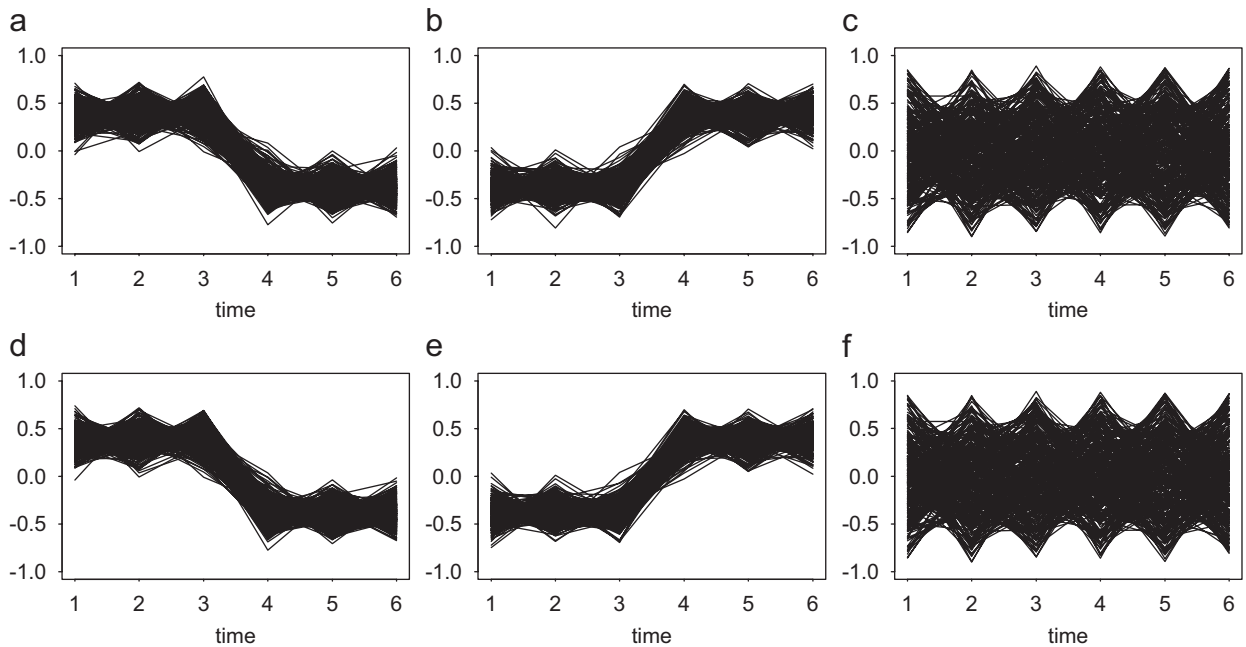


Fig. 3. A clustering result of the MG method on one transformed data set. (a)–(c): true clusters of the data. (d)–(f): Clusters obtained by the MG method.

Table 1  
Comparison of the clustering performance of the AHC, *K*-means, SOM, and MG methods for the 10 simulated data sets

Method	Untransformed data		Eigengenes		Transformed data	
	$\bar{\rho}$	SD	$\bar{\rho}$	SD	$\bar{\rho}$	SD
MG	0.9005	0.0048	—	—	0.9307	0.0027
AHC	0.5312	0.0069	0.5312	0.0069	0.8624	0.0118
<i>K</i> -means	0.7000	0.0102	0.7000	0.0102	0.8819	0.0043
SOM	0.8139	0.0081	0.7779	0.0066	0.9121	0.0029

$\bar{\rho}$ : the average of the adjusted Rand indices over the 10 data sets; SD: the standard deviation of the average.

(out of 1000) gene misclustered. The corresponding adjusted Rand index is  $\rho = 0.9338$ . Other computational results are summarized in Table 1. For comparison, the MG method was also applied to this example with the untransformed data. The BIC analysis suggests the VVV model for the untransformed data. The resulting adjusted Rand indices are summarized in Table 1.

For a thorough exploration for the effect of the SVD-based probit transformation, we also applied the agglomerative hierarchical clustering (AHC), *K*-means and SOM methods to this example with the untransformed data (*X*), the eigensamples (*U*), and the transformed data (*Z*). The softwares for AHC and *K*-means are available in S-plus. The software for SOM is developed by Tamayo et al. (1999) is downloadable at <http://www-genome.wi.mit.edu/software/genecluster2/gc2.html>. AHC was applied to each data set with the Euclidean distance and the average linkage. The dendrograms were cut such that the data were grouped into three clusters. The *K*-means method was run by restricting the number of clusters to be 3. In SOM, we set a  $1 \times 3$  grid for each data set, i.e., grouping the genes into three clusters. The other parameters were set to the default values as given in the software. The results are summarized in Table 1. The comparison shows that the probit transformation has significantly improved the performance of all clustering methods considered above for this example. The comparison also indicates that SVD alone cannot improve the performance of these clustering methods, and the probit transformation is crucial for this example.

In this example, all eigensamples were used in clustering. In practice, this is not necessary, as some eigensamples just represent the noise factor of the data. For most clustering methods, the clustering result based on an appropriate subset of eigensamples is often better than that based on all eigensamples. How to select eigensamples for use in clustering will be discussed in the next subsection.

### 2.3. Selection of eigensamples

In the literature of clustering, SVD is often used to reduce the dimensionality of the data by choosing only a subset of eigensamples to be used in clustering. Theoretically, the eigensamples can be regarded as projections of the observations on the directions of eigenpatterns. The contribution of the *i*th eigensample to the variation of the data can be measured by  $\xi_i = \sigma_i^2 / \sum_{j=1}^p \sigma_j^2$ ,  $i = 1, \dots, p$ . Based on this observation, many rules have been proposed for selection of eigensamples (Jolliffe, 1986). A common rule of thumb is to choose the first few eigensamples, as they contain most of the variations of the data. For example, we can set a threshold value  $\pi$ , which usually ranges between 0.7 and 0.9, and minimize the value of *m* such that the total contribution of the first *m* eigensamples to the variation of the data is no less than  $\pi$ , i.e., setting  $m = \operatorname{argmin}\{k : \sum_{i=1}^k \xi_i \geq \pi\}$ . However, Chang (1983) showed theoretically that the first few eigensamples may contain less cluster structure information than other eigensamples. He also constructed a two-component Gaussian mixture example, for which the two components are only well-separated in the subspace of the first and last eigensamples.

Since the contribution of each eigensample to the variation of the data is different, Kaiser (1960) suggests that if the contribution of some eigensample is too low, it is not worth retaining for clustering. Jolliffe (1986) suggests a cut-off value for  $\xi_i$ 's. If  $\xi_i < 0.7/p$ , then the *i*th eigensample can be discarded. However, this rule tends to retain too few eigensamples.

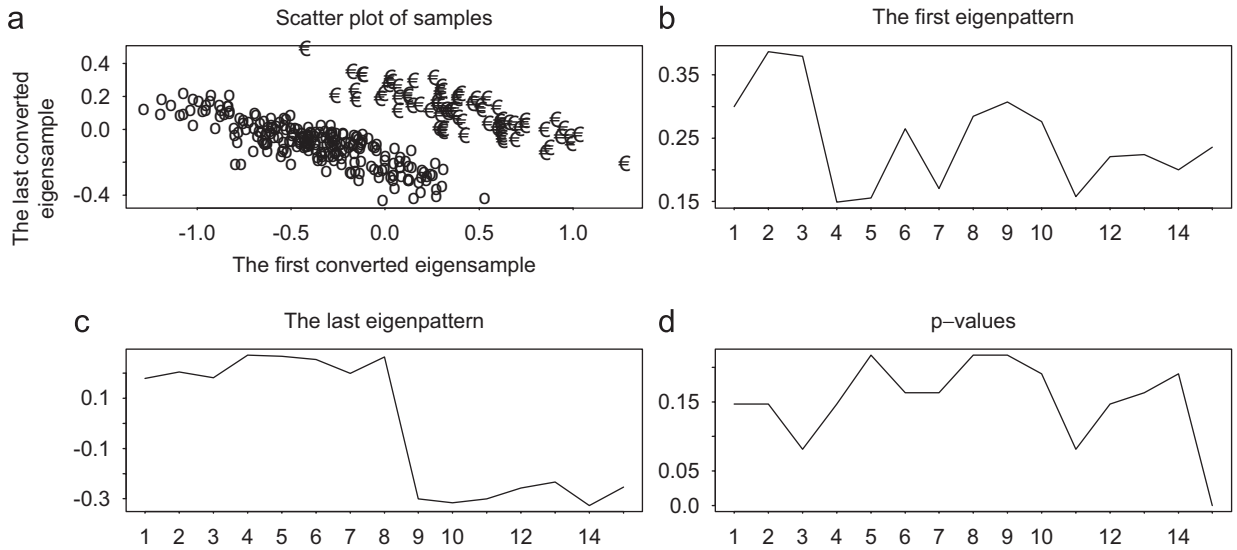


Fig. 4. SVD analysis for Chang’s example. (a) Scatter plot of the samples in the subspace of the transformed first and last eigensamples. (b) The first eigenpattern. (c) The last eigenpattern. (d) The  $p$ -values of the eigenpatterns.

In the following, we propose a new rule for selection of eigensamples. The new rule is a combination and modification of above two rules. The modification is made based on a run test for the randomness of eigenpatterns. The test can be described as follows.

- (a) Convert the eigenpattern into a binary sequence of zeros (for below median) and ones (for above median).
- (b) Let the random variable  $R$  denote the number of runs contained in the sequence. Calculate the corresponding  $p$ -value,

$$P(R = r) = \begin{cases} \left[ \binom{k_0 - 1}{s - 1} \binom{k_1 - 1}{s} + \binom{k_0 - 1}{s} \binom{k_1 - 1}{s - 1} \right] / \binom{k_0 + k_1}{k_0} & r \text{ is odd and } r = 2s + 1, \\ 2 \left[ \binom{k_0 - 1}{s - 1} \binom{k_1 - 1}{s - 1} \right] / \binom{k_0 + k_1}{k_0} & r \text{ is even and } r = 2s, \end{cases}$$

where  $k_0$  and  $k_1$  denote the number of zeros and the number of ones contained in the binary sequence, respectively.

If the  $p$ -value is lower than a specified significance level, the pattern is considered as significant; otherwise, it is considered as random. In our practice, 0.01 is often a good significance level for the run test. Refer to [Sprent and Smeeton \(2000\)](#) for more details on the test.

The run test can be illustrated by Chang’s example ([Chang, 1983](#)) as follows. The example consists of 300 observations of 15 dimensions. The observations are generated from the mixture Gaussian  $0.2N_{15}(\mu, \Sigma) + 0.8N_{15}(-\mu, \Sigma)$ , where the  $i$ th element of  $\mu$  is  $\mu_i = 0.95 - 0.05i$  for  $i = 1, 2, \dots, 15$ , the diagonal of elements of  $\Sigma$  is 1, and the off-diagonal elements of  $\Sigma$  are determined by the off-diagonal elements of the matrix  $-0.13(ff^T)$ . The first eight elements of  $f$  are  $-0.9$  and the last seven elements of  $f$  are  $0.5$ . [Fig. 4\(a\)](#) shows that the samples from the two components are well separated in the subspace of the transformed first and last eigensamples. [Figs. 5\(b\) and \(c\)](#) show the first and last eigenpatterns, respectively. [Fig. 4\(d\)](#) shows the  $p$ -values of the run tests for each of the eigenpatterns. The last eigenpattern is highly significant, and its  $p$ -value is  $3 \times 10^{-4}$ .

Motivated by Chang’s example, we suggest the following rule for selection of eigensamples. Let  $S$  denote the index set of the selected eigensamples,  $S \subset \{1, \dots, p\}$ . Let  $S_0$  denote the index set of the eigensamples with  $\xi_i$  greater than a cut-off value, say,  $0.2/p$ . Since the cut-off value we set here is low (lower than  $0.7/p$  the value suggested by [Jolliffe, 1986](#)),  $S_0$  will include more eigensamples, even some of them are not really worth retaining. In our experience, a cut-off value between  $0.1/p$  and  $0.3/p$  usually works well for the new rule described below. Let

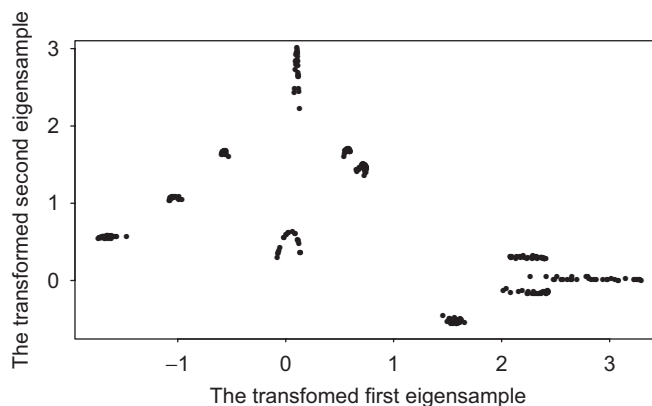


Fig. 5. Illustration of the SVD-based probit transformation for one data set generated in this example.

$S_1$  denote the index set of the significant eigenpatterns determined by the run test, and  $S_1^c$  be the complementary set of  $S_1$ .

- Determine the sets  $S_0$ ,  $S_1$  and  $S_1^c$ .
- Let  $S_2 = S_0 \cap S_1$  and  $S_2^* = S_0 \cap S_1^c$ . Arrange  $S_2$  and  $S_2^*$  in descending order of  $\xi_i$ 's.
- Set  $S = S_2$ . If  $\sum_{i \in S} \xi_i \geq \pi$ , go to step (d); otherwise, move eigensamples from  $S_2^*$  into  $S$  in the order given by  $S_2^*$  until  $\sum_{i \in S} \xi_i \geq \pi$  or  $S_2^*$  is empty.
- Convert the selected eigensamples into normal random variables using the probit transformation.

Henceforth, the new rule will be called the eigensample selection rule. If we apply the rule to Chang's example and set  $\pi$  to a number greater than 0.15, the first and last eigensamples will be included into the set  $S$ . The MG method can then produce a perfect clustering result with the adjusted Rand index  $\rho = 1$ . How to choose the value of  $\pi$  for a real data set will be discussed in Section 4.

For the transformed data, the correct number of clusters can also be identified according to the BIC criterion. However, if no probit transformation is used, the MG method will produce a clustering result with 38 observations being misclustered, and the corresponding adjusted Rand index is as low as 0.53. This shows again the probit transformation has the potential to improve the performance of the MG method.

### 3. Simulated examples

#### 3.1. Example 1

Yeung and Ruzzo (2001) considered a simulated example, which models the cyclic behavior of genes over different time points using the sine function. The sine function modeling for the cell cycle behavior is supported by the experiments reported by Holter et al. (2000) and Alter et al. (2000). Genes in the same cluster have similar peak time over the time course. Different clusters have different phase shifts and different sizes. This example is the same with Yeung and Ruzzo's example except that it consists of more genes and an extra cluster of scattered genes.

The example consists of 10 data sets, and each data set consists of 500 genes and 11 clusters. The cluster sizes are generated according to Zipf's law (Zipf, 1949), and the expression profiles are simulated as follows. Let  $s_i = (s_{i1}, \dots, s_{ip})$  denote the true expression profile of gene  $i$ , and let  $x_i = (x_{i1}, \dots, x_{ip})$  denote the observed one in the experiment, where  $p = 24$ . We set

$$s_{ij} = \lambda_j \left[ \alpha_i + \beta_{ki} \sin \left( \frac{2\pi j}{8} - w_k + \varepsilon \right) \right],$$

$$x_{ij} = s_{ij} + \delta_j, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

Table 2  
Summary of the computational results of the MG method for the simulated examples

Method	Untransformed		Transformed	
	$\bar{\rho}$	$\bar{G}$	$\bar{\rho}$	$\bar{G}$
Example I	0.7423 (0.0463)	6.9 (0.86)	0.8991 (0.0259)	9.2 (0.42)
Example II	0.8632 (0.0163)	11.0 (0.52)	0.9338 (0.010)	11.9 (0.31)
Example III	0.8983 (0.0237)	10.6 (0.4)	0.9837 (0.0044)	11.2 (0.13)

$\bar{\rho}$ : the average of adjusted Rand indices over 10 data sets.  $\bar{G}$ : the average of the number of clusters determined by the BIC criterion; the true number of clusters of the three examples are 11, 12 and 11, respectively. The numbers in the parentheses are the standard deviations of the corresponding averages.

where  $n = 500$ ,  $\delta_j \sim N(0, 1)$ ,  $\lambda_j \sim N(3, 0.5^2)$ ,  $\alpha_i \sim N(0, 1)$ ,  $\varepsilon \sim N(0, 1)$ ,  $w_k \sim \text{Unif}[0, 2\pi]$  for  $k = 1, \dots, 11$ ,  $\beta_{ki} \sim N(3, 0.5^2)$  for  $k = 1, \dots, 10$ , and  $\beta_{ki} = 0$  for  $k = 11$ . Here  $\alpha_i$  is the average expression level of gene  $i$ ,  $\beta_j$  is the amplitude control for gene  $i$ ,  $w_k$  is the phase shift,  $\varepsilon$  is the noise of gene synchronization,  $\lambda_j$  is the amplitude control for condition  $j$ , and  $\delta_j$  is an additive experiment error. Different clusters are represented by different  $w_k$ 's. The genes in cluster 11 are non-periodically expressed, and can be regarded as scattered genes. Each profile was then normalized to have mean 0 and unit length. For the non-scattered genes of this example, the average signal/noise ratio is about 6.7. Here the signal/noise ratio of gene  $i$  is defined as the ratio of the standard deviations of the signal vector  $s_i$  and the noise vector  $(\delta_1, \dots, \delta_p)$ . This example mimics the scenario that the scattered genes are present and the experimental error level is low.

The eigensample selection rule was applied to this example with  $\pi = 90\%$ . For simplicity, we fix  $\pi = 90\%$  for all three simulated examples studied in this section. For each data set of this example, only the first two eigensamples were retained for clustering. The probit transformation has reduced the dimensionality of the data greatly, from 23 to 2. The MG method was then applied to the transformed eigensamples with the covariance structure and the number of components being selected according to the BIC criterion. For comparison, the MG method was also applied to the untransformed data, and the covariance structure and the number of components of the model were also determined by the BIC criterion. The quality of the clustering results is assessed by the adjusted Rand index. The results are summarized in Table 2.

The comparison indicates that the SVD-based probit transformation has improved significantly the performance of the MG method for this example. The improvement can be drastic for some data sets. Fig. 5 illustrates the performance of the transformation for one data set generated in this example. For this data set, the eleven clusters are well separated in the subspace of the transformed first and second eigensamples. Applying the MG method to the transformed data leads to a perfect clustering result with the adjusted Rand index  $\rho = 1$ . However, when the MG method is applied to the untransformed data, the resulting adjusted Rand index is only 0.45.

### 3.2. Example II

This example consists of 10 data sets, and each data set consists of 1000 genes and 12 clusters. Let  $s^{(k)} = (s_1^{(k)}, \dots, s_p^{(k)})$  be the common expression pattern of the genes in cluster  $k$ , and let  $x_i^{(k)}$  be the observed expression profile of gene  $i$  in cluster  $k$ , where  $p = 15$  and  $k = 1, \dots, 12$ . The profiles were simulated as follows. Let  $\mathbf{a} = (-0.65, -0.55, \dots, 0.25, 0.25, 0.35, \dots, 0.65, 0.75)'$ ,  $\mathbf{b} = (0.25, 0.5, \dots, 1.25, 2.5, 2.75, \dots, 3.75)'$ ,  $\mathbf{d} = (1, \dots, p)'$ , and  $\mathbf{j}_p = (1, \dots, 1)'$ . For  $k = 1, \dots, 11$ , we set

$$\mathbf{s}^{(k)} = \psi(\cos(2a_k \mathbf{d} \pi / 5 + b_k \mathbf{j}_p)), \quad \mathbf{x}_i^{(k)} = \mathbf{s}^{(k)} + \boldsymbol{\epsilon}_i^{(k)}, \quad i = 1, \dots, n_k,$$

and for  $k = 12$ , we set

$$s^{(12)} = \mathbf{0}, \quad x_i^{(12)} = \epsilon_i^{(12)}, \quad i = 1, \dots, 1000 - \sum_{k=1}^{11} n_k,$$

where  $\cos(\mathbf{z}) = (\cos(z_1), \dots, \cos(z_p))$ ,  $n_k$  is the cluster size,  $\psi(\mathbf{z})$  denotes a normalization operator which normalizes  $\mathbf{z}$  to a vector with mean 0 and variance 1, and  $\epsilon_i^{(k)}$  is a random vector drawn from the multivariate normal distribution  $N_p(\mathbf{0}, \Sigma_k)$  with  $\Sigma_k$  being drawn from the inverse Wishart distribution  $IW(20, 0.25I_p)$ . The cluster sizes of the first 11 clusters were drawn from a Poisson distribution with mean 85, and they were selected such that  $n_{12} = 1000 - \sum_{i=1}^{11} n_i > 0$ . The genes in cluster 12 can be regarded as scattered genes. Each profile was then normalized to have mean 0 and unit length. For the non-scattered genes in each data set, the average signal/noise ratio is about 1.1. Here the signal/noise ratio of a gene expression profile, say,  $x_i^{(k)}$ , is defined as the ratio of the standard deviations of  $s^{(k)}$  and  $\epsilon_i^{(k)}$ . Hence, this example mimics the scenario that the scattered genes are present and the experiment error level is high. The MG method was applied to both the transformed and untransformed data. The results are summarized in Table 2.

### 3.3. Example III

This example consists of 10 data sets. Each data set consists of 1000 genes and 11 clusters. These clusters were generated in the same way as the first 11 clusters of Example II except that the cluster sizes are different. Here the sizes of the first 10 clusters were drawn from a Poisson distribution with mean 90, and they were selected such that  $n_{11} = 1000 - \sum_{k=1}^{10} n_k > 0$ . This example mimics the scenario that no scattered genes presented in the data set. The MG method was then applied to both the transformed and untransformed data. The computational results are summarized in Table 2.

Table 2 indicates that the SVD-based probit transformation has improved significantly the performance of the MG method for the above simulated examples. The transformation does not only improve the values of adjusted Rand indices, but also improves the estimates for the true number of clusters that were made according to the BIC criterion. These examples suggest that the use of the transformation can be quite general. The data sets can include or exclude scattered genes, and their experiment error levels can be high or low. In these examples,  $\pi$  is set to 0.9. Although this setting may not be optimal, it indicates for these examples the existence of a subset of eigensamples based on which the MG method can produce better clustering results than those based on the original data.

## 4. A real data example

The fibroblast data set was collected by Iyer et al. (1999) for the purpose of investigating the response of human fibroblast to serum after growth arrest. In the study, the temporal changes in mRNA levels of 8613 genes were measured at 12 time-points, ranging from 15 min to 24 h after serum stimulation. The full data set is available at <http://genome-www.stanford.edu/serunm/data.html>. The genes whose expression levels changed substantially in response to serum were extracted to form a subset of 517 genes. Iyer et al. (1999) analyzed this subset of data and suggested the existence of 11 clusters. In this paper, we reanalyzed this subset of data using the MG method.

Each profile was preprocessed by a logarithm transformation and then normalized to have mean 0 and unit length. The probit transformation was applied to the normalized data set. The value of  $\pi$ , the tuning parameter of the eigensample selection rule, can be determined according to the overall quality of the resulting clustering result. The overall quality of a clustering result can be measured using the DB index (Davies and Bouldin, 1979). Since the true clusters are unknown for real data, the adjusted Rand index used in simulated examples can no longer be used here. The DB index is a composite index reflecting a trade-off between the compactness and the separation of the clusters, and has been widely used to decide the optimal number of clusters in  $K$ -means. The index is defined as

$$DB = \frac{1}{G} \sum_{k=1}^G \max_{l \neq k} \left( \frac{S(C_k) + S(C_l)}{D(c_k, c_l)} \right),$$

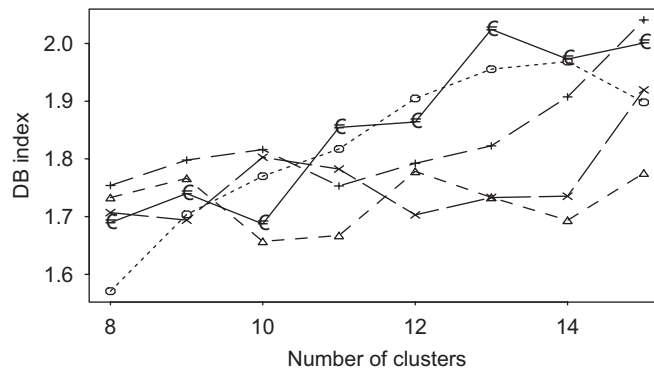


Fig. 6. DB indices produced by the MG method for the transformed and untransformed Fibroblast data. Black points (●): the DB indices for the untransformed data; circle (○): the DB indices for the transformed data with  $\pi = 0.85$ ; triangle (△): the DB indices for the transformed data with  $\pi = 0.9$ ; plus (+): the DB indices for the transformed data with  $\pi = 0.94$ ; times (×): the DB indices for the transformed data with  $\pi = 0.95$ .

where  $G$  is the number of clusters,  $S(C_k)$  is the average distance of all profiles in cluster  $k$  to their cluster center  $c_k$ , and  $D(c_k, c_l)$  is the distance between cluster centers  $c_k$  and  $c_l$ . A small value of the index indicates a better overall quality of the clustering result. Refer to Chen et al. (2002) for other composite indices of clustering evaluation.

Fig. 6 shows the DB indices of the clustering results produced by the MG method for the transformed data with different choices of  $\pi$  and different numbers of clusters. For comparison, the DB indices produced by the MG method for the untransformed data were also shown in the plot. The comparison indicates that the probit transformation can lead to better clustering results with an appropriate choice of  $\pi$ . The choice of  $\pi$  may depend on our choice for the number of clusters. For example, if we want to cluster the profiles into several big clusters, e.g., eight clusters, Fig. 6 suggests that the setting  $\pi = 0.85$  is appropriate. If we want to cluster the profiles into many small clusters, e.g., 11 or 14 clusters, Fig. 6 suggests that the setting  $\pi = 0.9$  is appropriate.

## 5. Conclusion

In this paper, we have proposed to use the SVD-based probit transformation to improve the performance of the MG method for clustering gene expression profiles. Our numerical results show that the transformation can be generally useful for both types of data sets with or without scattered genes. As a by-product, we show that the probit transformation also improves the performance of the model-free clustering methods, such as SOM, AHC and  $K$ -means, for the data sets containing scattered genes.

We have also proposed a run-test-based rule for selecting eigensamples used for clustering. The new rule works well for all examples studied in this paper. The new rule includes a tuning parameter  $\pi$ . In Section 4, we suggested to choose the value of  $\pi$  by a trial and error method: try different values of  $\pi$  and choose the one which leads to the best clustering result. Choice of  $\pi$  also reflects our belief on the signal/noise level of the data. If we think the data is quite noisy, we may set  $\pi$  to a small value. A small value of  $\pi$  can effectively exclude the noisy eigensamples from use in clustering. Otherwise, we may prefer to set  $\pi$  to a large value.

## Acknowledgments

The author's research was partially supported by grants from the National Science Foundation (DMS-0405748) and the National Cancer Institute (CA104620).

## References

- Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97, 10101–10106.
- Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821.

- Biernacki, C., Celeux, G., Govaert, G., 1999. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Lett.* 20, 267–272.
- Carr, D.B., Somogyi, R., Michaels, G., 1997. Templates for looking at gene expression clustering. *Statist. Comput. Statist. Graphics Newslett.* 8, 20–29.
- Chang, W.C., 1983. On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.* 32, 267–275.
- Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S., Zhang, M.Q., 2002. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statist. Sinica* 12, 241–262.
- Cho, R.J., Campbell, M.J., Winzler, E.A., Steinmetz, L., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W., 1998. A genome wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73.
- Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell. PAMI-1* (2), 224–227.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc., Ser. B* 39, 1–38.
- Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.
- Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* 97, 611–631.
- Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P., 2000. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 1 research0003.1-0003.21.
- Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., Fedoroff, N.V., 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA* 97, 8409–8414.
- Horn, D., Axel, I., 2003. Novel clustering algorithm for microarray expression data in a truncated SVD space. *Bioinformatics* 19, 1110–1115.
- Hubert, L., Arabie, P., 1985. Comparing partitions. *J. Classification*, 193–218.
- Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Satudt, L.M., Hudson Jr., J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., Brown, P.O., 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87.
- Jolliffe, I.T., 1986. *Principal Component Analysis*. Springer, New York.
- Kaiser, H.F., 1960. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20, 141–151.
- Liang, F., Wang, N., 2007. Dynamic agglomerative clustering of gene expression profiles. *Pattern Recogn. Lett.*, doi:10.1016/j.patrec.2007.01.009.
- McLachlan, G.J., Bean, R.W., Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18, 413–422.
- Medvedovic, M., Sivaganesan, S., 2002. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* 18, 1194–1206.
- Medvedovic, M., Yeung, K.Y., Bumgarner, R.E., 2004. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* 20, 1222–1232.
- Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* 66, 846–850.
- Sprent, P., Smeeton, N.C., 2000. *Appl. Nonparametric Statist. Methods*. third ed. Chapman & Hall/CRC, London.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96, 2907–2912.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285.
- Tou, J.T., Gonzalez, R.C., 1979. Pattern classification by distance functions. In: Tou, J.T., Gonzalez, R.C. (Eds.), *Pattern Recognition Principles*. Addison-Wesley, Reading, MA, pp. 75–109.
- Tseng, G.C., Wong, W.H., 2005. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics* 61, 10–16.
- Wakefield, J., Zhou, C., Self, S., 2003. Modeling gene expression over time: curve clustering with informative prior distributions. In: Bernardo, J.M., Bayarri, M.J., O, B.J., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), *Bayesian Statistics*, vol. 7. Clarendon Press, Oxford.
- Watkins, D.S., 1991. *Fundamentals of Matrix Computations*. Wiley, New York.
- Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L., 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987.
- Yeung, K.Y., Ruzzo, W.L., 2001. Principal component analysis for clustering gene expression data. *Bioinformatics* 17, 763–774.
- Zipf, G.K., 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, MA.