



CLUSTERING GENE EXPRESSION PROFILES USING MIXTURE MODEL ENSEMBLE AVERAGING APPROACH

FAMING LIANG

Department of Statistics
Texas A & M University
College Station, TX 77843-3143
U. S. A.
e-mail: fliang@stat.tamu.edu

Abstract

Clustering has been an important tool for extracting underlying gene expression patterns from massive microarray data. However, most of the existing clustering methods cannot automatically separate noise genes, including scattered, singleton and mini-cluster genes, from other genes. Inclusion of noise genes into regular clustering processes can impede identification of gene expression patterns. The model-based clustering method has the potential to automatically separate noise genes from other genes so that major gene expression patterns can be better identified. In this paper, we propose to use the ensemble averaging method to improve the performance of the single model-based clustering method. We also propose a new density estimator for noise genes for Gaussian mixture modeling. Our numerical results indicate that the ensemble averaging method outperforms other clustering methods, such as the quality-based method and the single model-based clustering method, in clustering datasets with noise genes.

2000 Mathematics Subject Classification: **Kindly Provide.**

Keywords and phrases: ensemble average, Gaussian mixture models, scatter genes, silhouette width.

Received April 16, 2008

1. Introduction

DNA microarray technology has made it possible to examine the expression of many genes over multiple developmental stages or under different experimental conditions. One of the important tasks is to identify co-expressed genes from messy DNA microarray data. Microarray experiments involve many noise genes, including scattered, singleton and mini-cluster genes. A gene is called a *scattered gene* if its expression level does not change much across samples. The scattered gene provides no or little information to the underlying biological processes, and shows low correlation to the expression patterns of non-scattered genes. Mini-cluster genes refer to the genes which belong to a very small cluster. A gene whose expression pattern is different from the expression patterns of any other genes is called a *singleton*. If the noise genes are forced into a cluster, the average profile of this cluster can be compromised and the composition of this cluster might provide less information for future analyses. As demonstrated by Tseng and Wong [28], the partitioning clustering methods, such as K-means, self-organizing maps (Kohonen [17]), and hierarchical methods, fail to produce reasonable clusters for a dataset containing noise genes.

In the literature, a common strategy to deal with scattered genes is to filter them away through a variation filter. The filter is usually set in a way such that only the genes bearing sufficient variation across different samples are retained for further analyses. For example, in clustering the yeast gene expression profiles (Cho et al. [6]), Tamayo et al. [27] filtered away the genes for which the relative expression change is less than 2 or the absolute expression change is less than 35 units across the samples. Due to the complexity of gene expression mechanism, it is impossible to set a filter which can avoid simultaneously the two types of errors, namely, removing some non-scattered genes from the dataset and leaving some scattered genes in the dataset. In addition, the filter is incapable of removing the singleton and mini-cluster genes.

An alternative strategy to deal with noise genes is sequential clustering. Included works are quality-based clustering (Heyer et al. [15]), adaptive quality-based clustering (AQC) (De Smet et al. [8]), CAST

(Ben-Dor et al. [2]), gene shaving (Hastie et al. [13]), CLICK (Sharan and Shamir [25]), HCS (Hartuv et al. [11]), tight clustering (Tseng and Wong [28]), Liang and Wang [18], among others. Taking AQC as an example, it first searches for a cluster center, and then groups the genes around the center into a cluster. Once a cluster is formed, the corresponding genes will be removed from the dataset and the process will be restarted for the remaining genes. A common drawback of the sequential clustering algorithms is that they involve some tuning parameters, which may be difficult to specify, even for an expert researcher in the context of clustering.

A promising method to deal with noise genes is model-based clustering, in which the data are typically modeled by a Gaussian mixture distribution with the noise observations being handled by adding an extra component. By assuming that the noise observations are uniformly distributed in the data region, Banfield and Raftery [1], Dasgupta and Raftery [7], Campbell et al. [3, 4], McLachlan and Peel [20], and Fraley and Raftery [10] modeled the data by the following mixture distribution,

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{\tau_0}{V} + \sum_{k=1}^g \tau_k \phi_k(\mathbf{x}|\boldsymbol{\theta}_k), \quad (1)$$

where V is the hypervolume of the data region, g is the number of clusters, $\tau_k \geq 0$, $\sum_{k=0}^g \tau_k = 1$, $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ contains the parameters of component k , $\Theta = (\tau_0, \dots, \tau_g, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)$ contains all parameters of the model, and ϕ_k is a multivariate Gaussian density, i.e.,

$$\phi_k(\mathbf{x}|\boldsymbol{\theta}_k) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}. \quad (2)$$

In Fraley and Raftery [10], the hypervolume V is taken to be the volume of the smallest hyperrectangle with sides parallel to the axes that contains all of the data points, the parameters of the model are estimated using the EM algorithm (Dempster et al. [9]), and the number of clusters and the structure of the covariance matrices are determined according to the BIC criterion. A significant advantage of the model-based clustering method is that it is free of tuning parameters.

On the other hand, the ensemble averaging method (Wolpert [29]; Perrone [22]; Hashem [12]) has often been used in the context of supervised learning to improve the accuracy of prediction. Let \mathbf{x} denote a new input vector of a model, let y denote the corresponding response, and let $S_1(\mathbf{x}), \dots, S_L(\mathbf{x})$ denote L input-output functions realized by the model based on the same training data. These input-output functions can be varied on various aspects, such as model structure and training conditions. The ensemble averaging method suggests to use a combined predictor

$$S_{ea}(\mathbf{x}) = \sum_{i=1}^L \lambda_i S_i(\mathbf{x}),$$

to predict $E(Y | \mathbf{X} = \mathbf{x})$, where λ_i is called the *ensemble weight* of $S_i(\mathbf{x})$, $0 \leq \lambda_i \leq 1$, and $\sum_{i=1}^L \lambda_i = 1$. The choice of λ can depend on the predictors $S_1(\mathbf{x}), \dots, S_L(\mathbf{x})$. For example, if $S_1(\mathbf{x}), \dots, S_L(\mathbf{x})$ have the same model structure and are trained under different initial conditions, we can simply set $\lambda_i = 1/L$ for $i = 1, \dots, L$, the resulting predictor $S_{ea}(\mathbf{x})$ will have a smaller variance than any $S_i(\mathbf{x})$. Another use of the ensemble averaging method is to reduce the prediction bias caused by underfitting or overfitting by averaging over different models. Refer to Haykin [14, Chapter 7] for more discussions on the method.

The label switching method (Stephens [26]) used in the context of Bayesian clustering can be regarded as a cluster ensemble averaging method, but it only works for the models which have the same number of clusters. The consensus clustering method (Monti et al. [21]) works in the spirit of ensemble averaging, but the averaging is taken over the clustering results for different subsets of genes. In this paper, we show that the ensemble averaging method can significantly improve the clustering accuracy of the single model-based clustering method. In this paper, we also propose a robust density estimator for noise genes.

The remainder of this paper is organized as follows. In Section 2, we describe the ensemble averaging method for clustering normalized gene expression profiles, as a summary of the pioneering work (Stephens [26];

Monti et al. [21]) in this direction. In this section, we also propose a robust density estimator for noise genes. In Section 3, we illustrate the new method using two simulated examples. In Section 4, we apply the ensemble averaging method to two real gene expression datasets. In Section 5, we conclude the paper with a brief discussion.

2. Clustering Normalized Gene Expression Profiles

2.1. Ensemble averaging for clustering

Let $\mathbf{c}^{(g,k)} = (c_1^{(g,k)}, \dots, c_n^{(g,k)})$ denote a cluster assignment of the genes, where n is the number of genes in the dataset, g indicates a model with g clusters, and k indicates that the assignment was obtained at the k th training of the model g . Let $\mathbf{D} = (d_{ij})$ denote an $n \times n$ distance matrix. A general cluster ensemble averaging method can be described as follows.

(a) Initialize $\mathbf{D} = \mathbf{0}$; specify the model range $G_{\min} \leq g \leq G_{\max}$ and the number of model training times K ; and determine the ensemble weights λ_{gk} for $g = G_{\min}, \dots, G_{\max}$ and $k = 1, \dots, K$. Here G_{\min} and G_{\max} are called the *minimum and maximum numbers of clusters*, respectively.

(b) Train each model K times independently to produce the cluster assignments $\mathbf{c}^{(g,k)}$, $g = G_{\min}, \dots, G_{\max}$ and $k = 1, \dots, K$.

(c) Set $d_{ij} = \sum_{g=G_{\min}}^{G_{\max}} \sum_{k=1}^K \lambda_{gk} (1 - \delta(c_i^{(g,k)} = c_j^{(g,k)}))$ for $i, j = 1, \dots, n$,

where $\delta(\cdot)$ is the indicator function.

(d) Recluster genes using the hierarchical clustering method according to the distance matrix \mathbf{D} .

Through the construction of the distance matrix \mathbf{D} , the cluster information carried by different models are combined together. Since the models may have different number of clusters, \mathbf{D} provides a hierarchical distance measure for the clustership of genes. In this paper, the Gaussian mixture models are assumed for the gene expression profiles, and the models are trained using the EM algorithm as described in Section 2.2.

The ensemble weights are selected to be the same for each model, i.e., $\lambda_{gk} = 1/[K(G_{\max} - G_{\min} + 1)]$ for all g and k . A more sophisticated setting of λ_{gk} is to link its value with an overall quality measure of the corresponding clusters, for example, the BIC value of the model.

The number of clusters to be grouped finally in step (d) can be determined according to an overall quality measure of the resulting clusters, e.g., the average silhouette width (Rousseeuw [24]). Refer to Chen et al. [5] for other possible measures. The average silhouette width is a composite index reflecting the compactness and separation of clusters. For each gene i , its silhouette width is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (3)$$

where $a(i)$ is the average distance of gene i to other genes in the same cluster, and $b(i)$ is the average distance of gene i in its nearest neighbor cluster. In the context of ensemble averaging, d_{ij} is referred to as the distance between gene i and gene j . The average silhouette width $\bar{s} = \sum_{i=1}^n s(i)/n$ provides a measure for the overall quality of a clustering result. The larger value of \bar{s} is, the better is the overall quality of the clustering result. According to this criterion, we set the best number of clusters to be

$$G = \arg \max_{G'_{\min} \leq g \leq G'_{\max}} \bar{s}^{(g)}, \quad (4)$$

where $\bar{s}^{(g)}$ denotes the resulting average silhouette width if the data are grouped into g clusters. Note that G'_{\min} and G'_{\max} are not necessarily the same with G_{\min} and G_{\max} , although they are set to be so in this paper.

2.2. Model training

In this subsection, we first give a new density estimator for noise genes and then describe how to train a Gaussian mixture model using the EM algorithm. To get a theoretical explanation for the new density estimator, we prove the following theorem.

Theorem 2.1. Let $\mathbf{z} = (z_1, \dots, z_p)'$ denote a real p -vector, let $\mathbf{j}_p = (1, \dots, 1)'$ be a p -vector of ones, and let I_p be a $p \times p$ identity matrix. Define $\bar{z} = \frac{1}{p} \sum_{i=1}^n z_i$, $s^2 = \frac{1}{p-1} \sum_{i=1}^p (z_i - \bar{z})^2$, and $\tilde{\mathbf{z}} = (z_1, \dots, z_{p-1})' - \bar{z} \mathbf{j}_{p-1}$. If the variables z_1, \dots, z_p are mutually independent and have the common mean μ and the common variance σ^2 , then the following results hold,

$$(a) \text{Cov}(\mathbf{z} - \bar{z} \mathbf{j}_p) = \sigma^2 \left(I_p - \frac{1}{p} \mathbf{j}_p \mathbf{j}_p' \right).$$

$$(b) \tilde{\mathbf{z}}' \left(I_{p-1} - \frac{1}{p} \mathbf{j}_{p-1} \mathbf{j}_{p-1}' \right)^{-1} \tilde{\mathbf{z}} / \sigma^2 = (p-1) s^2 / \sigma^2.$$

Proof. (a) It follows from the results that $\text{Var}(x_i - \bar{x}) = \sigma^2 \left(1 - \frac{1}{p} \right)$ and that $\text{Cov}(x_i - \bar{x}, x_j - \bar{x}) = -\sigma^2 / n$.

(b) By noting that $\left(I_{p-1} - \frac{1}{p} \mathbf{j}_{p-1} \mathbf{j}_{p-1}' \right)^{-1} = I_{p-1} + \mathbf{j}_{p-1} \mathbf{j}_{p-1}'$, it can be verified directly that $\tilde{\mathbf{z}}' (I_{p-1} + \mathbf{j}_{p-1} \mathbf{j}_{p-1}') \tilde{\mathbf{z}} = (p-1) s^2$.

From a biological point of view, we are primarily interested in the relative up/down-regulation of gene expressions instead of the absolute amplitude changes. Hence, the gene expression profiles are usually normalized to have mean 0 and variance 1 before clustering. Let $\mathbf{X} = (x_{ij})$, $i = 1, \dots, n$ and $j = 1, \dots, p$, denote the normalized gene expression profile matrix. Each row of \mathbf{X} represents the expression profile of a gene. Since each row of \mathbf{X} has sum 0, we only need to work on $\tilde{\mathbf{X}}$, a submatrix comprising any $p-1$ columns of \mathbf{X} . In the following, we will use $\tilde{\mathbf{x}}$, a row of $\tilde{\mathbf{X}}$, to denote the expression profile of a gene.

Since noise genes do not represent any significant patterns, we can generally assume that the noise genes satisfy the conditions of Theorem 2.1; that is, the expression levels measured at different experiments are mutually independent and have a common mean and a common variance.

By assuming the normality of gene expression data and the validity of the variance approximation $s^2 \approx \sigma^2$, it follows from Theorem 2.1 that each noise gene bears the following density value,

$$\phi_0(\tilde{\mathbf{x}}) \approx \frac{1}{(2\pi)^{(p-1)/2} \left| I_{p-1} - \frac{1}{p} \mathbf{J}_{p-1} \mathbf{J}'_{p-1} \right|^{1/2}} e^{-\frac{p-1}{2}} = \frac{p^{1/2}}{(2\pi e)^{(p-1)/2}}. \quad (5)$$

Note that $f(\tilde{\mathbf{x}})$ does not only provide a density estimate for noise genes, it also provides a threshold value for non-noise clusters. If the model is trained under the maximum likelihood principle, any genes in a cluster with density values less than $f(\tilde{\mathbf{x}})$ should be grouped as noise genes. Hence, the new estimate enforces the production of neat clusters.

Since $\sqrt{2\pi e} \approx 4$, the new estimator (5) is approximately equivalent to the one given by Fraley and Raftery [10] for normalized gene expression profiles. In Fraley and Raftery [10], the noise are assumed to be uniformly distributed on the smallest hyperrectangle with sides parallel to the axes that contains all of the data points, and thus $V \approx 4^{p-1}$. However, our estimator is more robust to noise gene expression profiles because of its independence on the real observations.

With the new density estimator (5), the parameters of the model (1) can be estimated using the EM algorithm (Dempster et al. [9]) as follows.

- **E-step.** Denote the current parameter values as Θ , and calculate

$$\omega_{ik} = \frac{\tau_k \phi_k(\tilde{\mathbf{x}}_i | \boldsymbol{\theta}_k)}{\sum_{k=0}^G \tau_k \phi_k(\tilde{\mathbf{x}}_i | \boldsymbol{\theta}_k)}, \quad i = 1, \dots, n, \quad k = 0, \dots, g,$$

where, to simplify the expression, we let $\boldsymbol{\theta}_0 = \emptyset$ and $\phi_0(\tilde{\mathbf{x}}_i | \boldsymbol{\theta}_0) = \phi_0(\tilde{\mathbf{x}}_i)$.

- **M-step.** Update the parameter values by setting

$$\hat{\tau}_k = \frac{1}{n} \sum_{i=1}^n \omega_{ik}; \quad \hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^n \omega_{ik} \tilde{\mathbf{x}}_i}{\sum_{i=1}^n \omega_{ik}}; \quad \hat{\Sigma}_k = \frac{\sum_{i=1}^n \omega_{ik} (\tilde{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_k) (\tilde{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_k)'}{\sum_{i=1}^n \omega_{ik}}.$$

The algorithm is initialized by selecting g random gene expression profiles as initial means, selecting the identity matrix for each of g initial covariance matrices, and selecting each cluster weighted equally. The algorithm is iterated until the likelihood value is not changed in a significant manner from one iteration to the next. Refer to Little and Rubin [19] for more discussions on the EM algorithm.

3. Illustrative Examples

3.1. Example I

This example tests the performance of the ensemble averaging method in presence of singleton and mini-cluster genes. It mimics the case that a high cutoff value is set for the filter such that all scattered genes have been removed from the dataset. This example consists of 10 datasets, and each dataset consists of 500 genes generated as follows. Let $\mathbf{x}_i^{(k)} = (x_{i1}^{(k)}, \dots, x_{ip}^{(k)})$ be the simulated expression profile of gene i of cluster k , where $p = 15$ and $k = 1, \dots, 11$. 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 $\mathbf{j}_p = (1, \dots, 1)'$. Set

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

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 $\boldsymbol{\varepsilon}_i^{(k)}$ is a random vector drawn from the multivariate normal distribution $N_p(\mathbf{0}, \Sigma_k)$ with Σ_k being drawn from the inverse Wishart distribution $IW(20, 0.25I_p)$. The cluster sizes n_1, \dots, n_5 are drawn from a Poisson distribution with mean 3, and the cluster sizes n_6, \dots, n_{10} are drawn from a Poisson distribution with mean 80. The cluster sizes are selected such that $n_{11} = 500 - \sum_{i=1}^{10} n_i > 0$. Since n_1, \dots, n_5 are usually very small comparing to n_6, \dots, n_{11} , the genes in groups 1-5 can be treated as noise genes in clustering.

Since the true clusters of the simulated genes are known, we follow Yeung and Ruzzo [31] and Yeung et al. [30] to use the adjusted Rand index (Rand [23]; Hubert and Arabie [16]) to assess the overall quality of the clustering results. The adjusted Rand index measures the degree of agreement between two partitions of the same set of observations, even when the comparing partitions having different numbers of clusters. Let Ω denote a set of n observations, let $C = \{c_1, \dots, c_s\}$ and $C' = \{c'_1, \dots, c'_t\}$ represent two partitions of Ω , let n_{ij} be the number of observations that are in both cluster c_i and cluster c'_j , let $n_{i\cdot}$ be the number of observations in cluster c_i , and let $n_{\cdot j}$ be the number of observations in cluster c'_j . The adjusted Rand index is defined as

$$\rho = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \right] / \binom{n}{2}}{\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} \right] / 2 - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \right] / \binom{n}{2}}. \quad (6)$$

A higher value of ρ means a higher correspondence between the two partitions. When the two partitions are identical, ρ is 1. When a partition is random, the expectation of ρ is 0.

The ensemble averaging method was first applied to the datasets with the minimum number of clusters $G_{\min} = 5$ and the maximum number of clusters $G_{\max} = 10$. For each $g \in \{G_{\min}, \dots, G_{\max}\}$, the model was trained 5 times independently with different initializations. Figure 1 shows the clusters obtained for one dataset. It indicates that the ensemble averaging method has successfully identified all true clusters and grouped the noise genes into the noise cluster. Table 1 compares the adjusted Rand indices of the clusters produced by the ensemble averaging method and those produced by the single model-based clustering method. Note that the clusters produced by the single model-based clustering method form the ensemble that the averaging is taken. The comparison indicates that the ensemble averaging method has made significant improvement over the single model-based clustering method for this example. For example, when the number of clusters is restricted to 7, the average of adjusted Rand indices over all 50 models (10 datasets

$\times 5$ runs) is only 0.665, and the average of adjusted Rand indices over the 10 best BIC models (the model with the best BIC value among those produced in the 5 runs is called the *best BIC model*) is only 0.807, while the adjusted Rand index of the clustering result produced by the ensemble averaging method is 0.975. The ensemble averaging method can produce so accurate clustering results, because it combines clustering information from different models.

Figure 2(a) shows the average silhouette widths of the clusters produced by the ensemble averaging method for the 10 datasets. For 9 datasets, the correct number of clusters can be determined by maximizing the average silhouette width of the resulting clusters. For comparison, we also plot in Figure 2(b) the best BIC values produced by the EM algorithm for the 10 datasets. For only 3 datasets, the correct number of

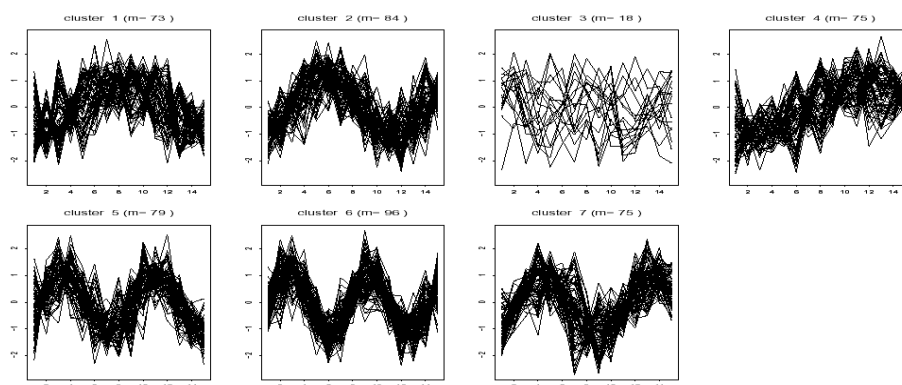


Figure 1. Clusters obtained by the ensemble averaging method for one dataset of Example I. The adjusted Rand index of the clusters is $\rho=0.994$. Cluster 3 corresponds to the noise cluster. The true number of noise genes is 16.

Table 1. Adjusted Rand indices of the clustering results for Example I. a: Average of adjusted Rand indices over 10 datasets; b: standard deviation of the corresponding average; c: average of adjusted Rand indices over 10 datasets \times 5 runs; d: average of adjusted Rand indices over the best BIC models (among the five runs) of the 10 datasets

Number of Clusters	5	6	7	8	9	10
Ensemble method ^a	0.709	0.841	0.975	0.969	0.964	0.960
Standard deviation ^b	0.014	0.017	0.005	0.004	0.005	0.006
All models in ensemble ^c	0.508	0.631	0.665	0.737	0.737	0.760
Standard deviation ^b	0.025	0.034	0.020	0.026	0.018	0.009
Best BIC models ^d	0.632	0.733	0.807	0.847	0.852	0.828
Standard deviation ^b	0.030	0.033	0.039	0.034	0.021	0.029

clusters can be determined by the BIC criterion.

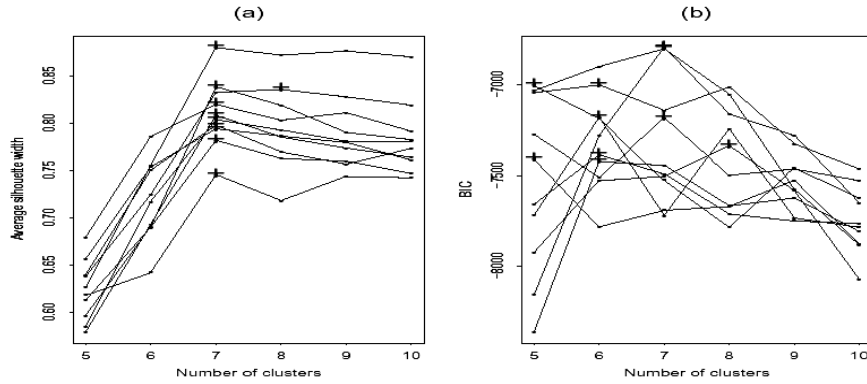


Figure 2. (a) Average silhouette width of the clusters produced by the ensemble averaging method for 10 datasets of Example I. The symbol ‘+’ indicates the number of clusters determined by the criterion (4). (b) Best BIC values (among five runs) produced by the EM algorithm for the 10 datasets. The symbol ‘+’ indicates the number of clusters determined by the BIC criterion.

For comparison, AQC was also applied to the dataset shown in Figure 1. The software for AQC was developed by De Smet et al. [8] and is

available at

<http://www.esat.kuleuven.ac.be/~thijs/Work/Clustering.html>.

For this dataset, we set the minimum cluster size $n_s = 30$ and the test significance level $q_s = 0.95$. Figure 3 shows the resulting clusters. AQC cannot identify all patterns contained in this dataset under the above setting, and it leaves too many genes unclustered. Other parameter settings have been tried, for example, $n_s = 30$ and $q_s = 0.9$, $n_s = 30$ and $q_s = 0.85$, and $n_s = 40$ and $q_s = 0.85$, the resulting clusters are similar. AQC tends to identify too many genes as noise genes. The true number of noise genes of this dataset is 16, while the number of noise genes identified by AQC under the above setting is 198.

3.2. Example II

This example tests the performance of the ensemble averaging method in presence of singleton, mincluster and clustered genes. It mimics the case that a low cutoff value is set for the filter such that some scattered genes are still left in the dataset. This example consists of 10 datasets. Each dataset consists of 1000 genes and 12 clusters. The first 11 clusters are generated as in Example I except that the cluster sizes are different. In this example, n_1, \dots, n_{10} are drawn from a Poisson distribution with mean 90, and n_{11} is drawn from a Poisson distribution

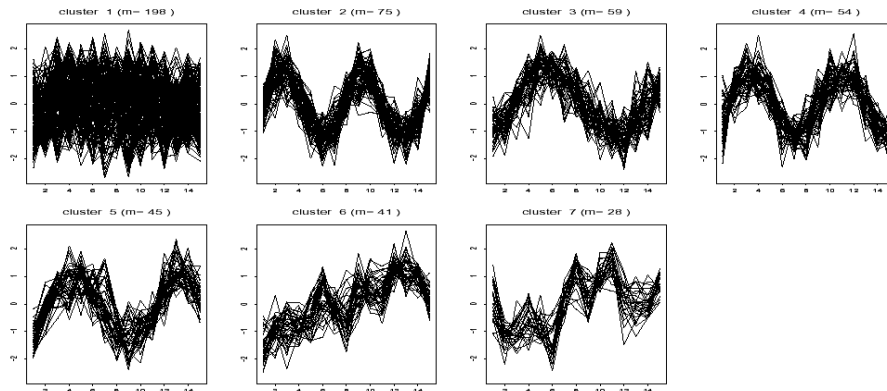


Figure 3. Clusters obtained by AQC for one dataset of Example I. Cluster 1 corresponds to the noise cluster. The true number of noise genes is 16.

with mean 3. The cluster sizes are selected such that $\sum_{i=1}^{11} n_i < 1000$.

The gene expression profiles in cluster 12 are generated as follows:

$$\mathbf{x}_i^{(12)} = \psi(\mathbf{u}_i), \quad i = 1, \dots, 1000 - \sum_{i=1}^{11} n_i,$$

where $\mathbf{u}_i = (u_{i1}, \dots, u_{ip})$ is a p -vector with u_{ip} being drawn independently from the uniform distribution $\text{Unif}[-2, 2]$. The genes in clusters 11 and 12 can be regarded as mini-cluster genes and scattered genes, respectively; and they together form the noise cluster of this example.

The ensemble averaging method was applied to each dataset with $G_{\min} = 8$ and $G_{\max} = 15$. For each $g \in \{G_{\min}, \dots, G_{\max}\}$, the model was trained 5 times independently with different initializations. Figure 4 shows the clusters obtained for one dataset. Table 2 compares the adjusted Rand indices of the clusters produced by the ensemble averaging method and those produced by the single model-based clustering method. Figure 5 compares the criterion (4) and the BIC criterion for determining the number of clusters for this example. The results indicate that the ensemble averaging method works well for this example even in presence of scattered genes, and the criterion (4) works better than the BIC criterion for determining the number of clusters.

For comparison, AQC was also applied to the dataset shown in Figure 4 with the parameters $n_s = 30$ and $q_s = 0.95$. The resulting clusters are shown in Figure 6. As in Example I, AQC fails to identify all patterns contained in the dataset, and groups too many genes as noise genes. Different parameter settings are tried, for example, $n_s = 30$ and $q_s = 0.9$, $n_s = 30$ and $q_s = 0.85$, and $n_s = 40$, and $q_s = 0.85$, the results are similar.

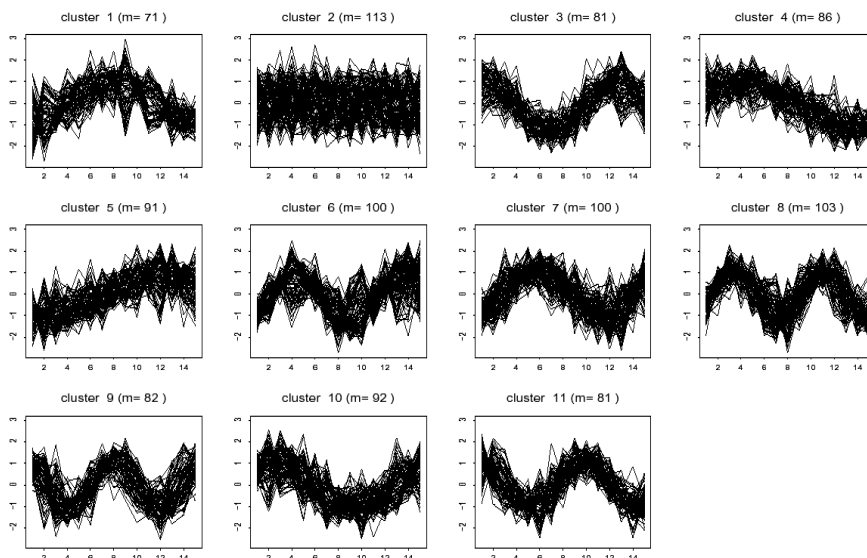


Figure 4. Clusters obtained by the ensemble averaging method for one dataset of Example II. The adjusted Rand index of the clusters is $\rho=0.971$. Cluster 2 corresponds to the noise cluster. The true number of noise genes is 104.

Table 2. Adjusted Rand indices of the clustering results for Example II. Refer to Table 1 for the notations of the table

Number of Clusters	8	9	10	11	12	13	14	15
Ensemble method	0.733	0.799	0.882	0.969	0.971	0.970	0.970	0.969
Standard deviation	0.005	0.006	0.007	0.007	0.006	0.006	0.006	0.005
All models in ensemble	0.622	0.660	0.701	0.758	0.770	0.770	0.807	0.795
Standard deviation	0.016	0.013	0.016	0.013	0.018	0.011	0.017	0.018
Best BIC models	0.690	0.735	0.813	0.845	0.850	0.891	0.892	0.884
Standard deviation	0.017	0.016	0.025	0.013	0.024	0.017	0.017	0.015

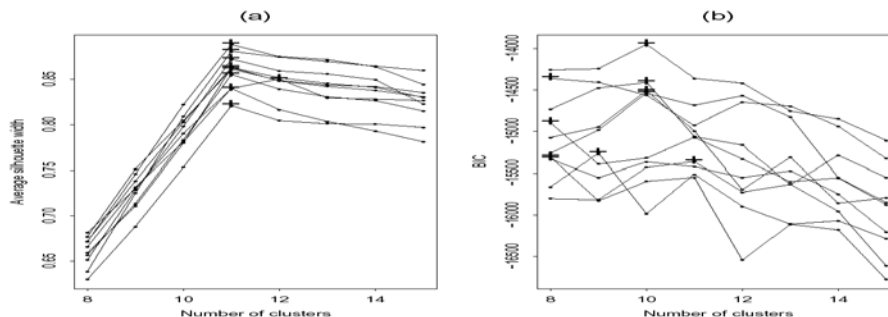


Figure 5. (a) Average silhouette width of the clusters produced by the ensemble averaging method for the 10 datasets of Example II. The symbol ‘+’ indicates the number of clusters determined by the criterion (4). (b) Best BIC values (among five runs) produced by the EM algorithm for the 10 datasets. The symbol ‘+’ indicates the number of clusters determined by the BIC criterion.

4. Real Data Examples

4.1. Leukemia cell line HL-60 data

The myeloid leukemia cell line HL-60 undergoes macrophage differentiation on treatment with the phorbol ester PMA. Nearly 100% of HL-60 cells become adherent and exit the cell cycle with 24 hours of PMA treatment. To monitor the process, expression levels of more than 6000 human genes were measured at four time points 0, 0.5, 4 and 24 hours after PMA stimulation. This dataset is available at

<http://www-genome.wi.mit.edu/software/genecluster2/gc2.html>

and has been used by Tamayo et al. [27] as an example to support the use of SOM.

In this paper, we use this dataset to demonstrate that the ensemble averaging method can make a further improvement in a dataset with scattered genes being removed in advance by a variation filter. In particular, the ensemble averaging method can automatically separate the singleton and mini-cluster genes from other genes so that the patterns in tighter large clusters can be better identified. The variation filter used here is the same as that used in Tamayo et al. [27]. Totally there are 590 genes left after filtration. This number is slightly different

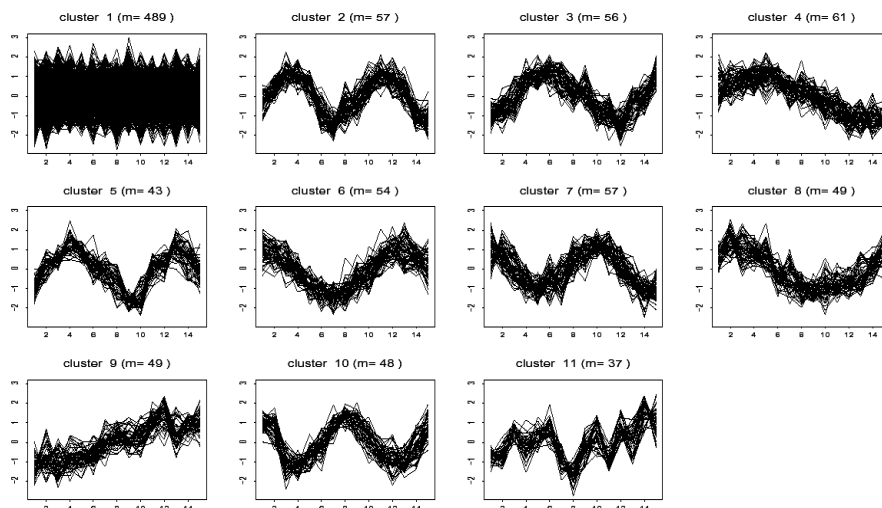


Figure 6. Clusters obtained by AQC for one dataset of Example I. Cluster 1 corresponds to the noise cluster. The true number of noise genes is 104.

from the number (567) reported in Tamayo et al. [27]. The expression profiles of the 590 genes were then normalized such that each has mean 0 and variance 1.

The ensemble averaging method was first applied to this pre-processed data with $G_{\min} = 5$ and $G_{\max} = 15$. For each $g \in \{G_{\min}, \dots, G_{\max}\}$, the EM algorithm was run 20 times independently. Figure 7 (left panel) shows that $g = 12$ is appropriate for this dataset. Figure 8 shows the resulting 12 clusters. Note that clusters 10 and 11 represent similar patterns. This is quite typical for the model-based clustering methods, which often split a big cluster into several similar small clusters so that the total likelihood value of the model is increased. The same phenomenon can also be observed in Figures 11 and 12 for the LD example. Clearly this is not a big drawback of the mixture model-based methods, because it does not hinder the finding of major gene expression patterns from the messy dataset.

For comparison, SOM and AQC were also applied to the pre-processed data. In SOM clustering, we follow Tamayo et al. [27] to specify

a grid of size 4×3 for the dataset. The resulting 12 clusters are shown in Figure 9. SOM performs very well for this example. It is only in its comparison to outcomes of the ensemble averaging method that we see that the outcome could be further improved by removing singleton and mini-cluster genes.

AQC was applied to the pre-processed data with the minimum cluster size $n_s = 2$ and the test significance level $q_s = 0.725$. AQC finds 11 clusters which are shown in Figure 10. Here we set a very low test

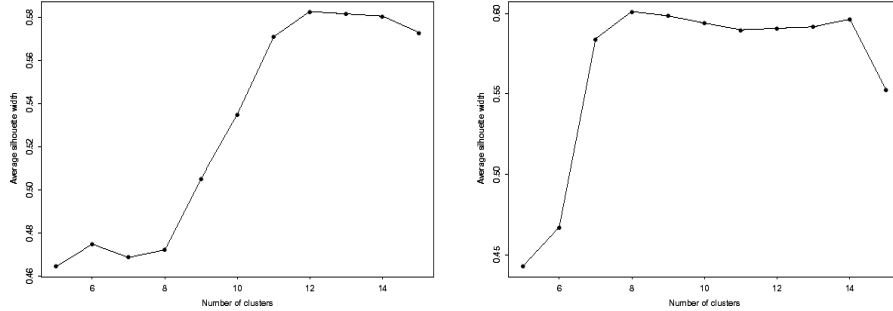


Figure 7. Left: Average silhouette width of the clusters produced by the ensemble averaging method for the leukemia cell line HL-60 example. Right: Average silhouette width of the clusters produced by the ensemble averaging method for the LD example.

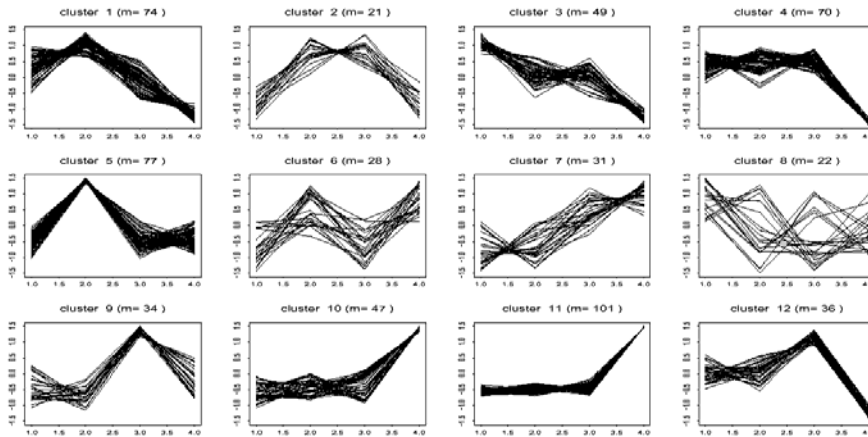


Figure 8. Clustering result of the ensemble averaging method for the pre-processed HL-60 data. Cluster 8 is the noise cluster.

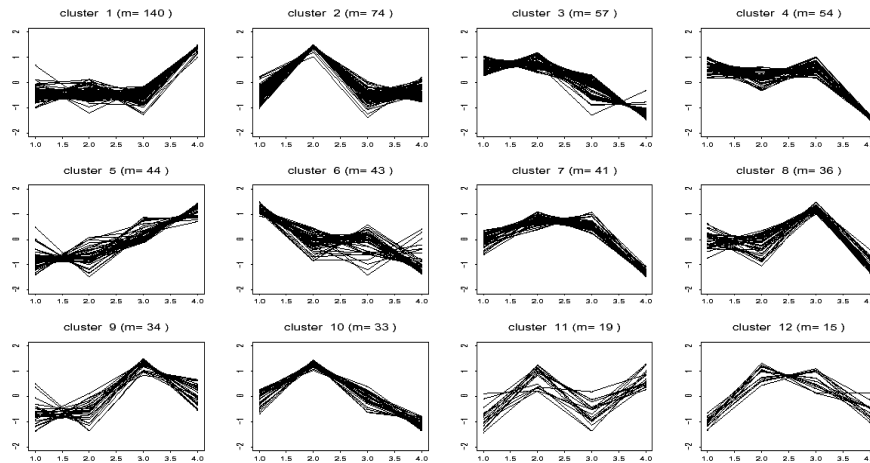


Figure 9. Clustering result of SOM for the pre-processed HL-60 data.

significance level, because AQC finds no clusters for this example when the test significance level is greater than 0.85. Figure 10 indicates that AQC tends to produce tight clusters and leave too many genes unclustered. This finding is consistent with the results presented in Figures 3 and 6 for the simulated examples.

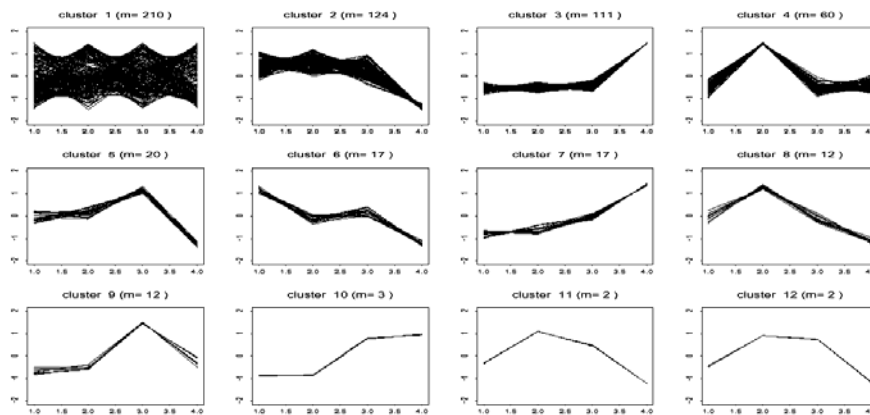


Figure 10. Clustering result of AQC for the pre-processed HL-60 data. Cluster 1: collection of the genes unclustered by AQC.

4.2. Avian pineal gland gene expression data

The avian pineal gland contains both circadian oscillators and photoreceptors to produce rhythms in biosynthesis of the hormone

melatonin in vivo and in vitro. It is of great interest to understand the genetic mechanisms driving the rhythms. For this purpose, a sequence of cDNA microarrays of birds' pineal gland transcripts under the light-dark (LD) condition was generated. The birds were euthanized at 2, 6, 10, 14, 18, 22 hours Zeitgeber time (ZT) to obtain mRNA to produce adequate cDNA libraries. Four microarray chips per time point were produced. Throughout the experiment, samples from LD ZT18 were used as controls. Four observations at each time point were log-transformed and averaged. This produces a data matrix of size 7730×6 . Each row represents the expression profile of a gene at six time points. The filter was set to remove the genes for which the variance of the expression levels at six time points is less than 1.0. After the filtration, 780 genes were left. Their expression profiles were then normalized to have mean 0 and variance 1.

The ensemble averaging method was applied to the pre-processed dataset with $G_{\min} = 5$ and $G_{\max} = 15$. For each $g \in \{G_{\min}, \dots, G_{\max}\}$, the EM algorithm was run 20 times. Figure 7 (right panel) shows that $g = 8$ is appropriate for the dataset. Figure 11 shows the resulting clusters. Five major gene expression patterns are identified from the dataset, and they are represented by clusters 2, 3, 4, 5, and 8, respectively. Here clusters 3, 6 and 7 represent similar patterns. As explained before, this is a typical phenomenon for the model-based clustering method.

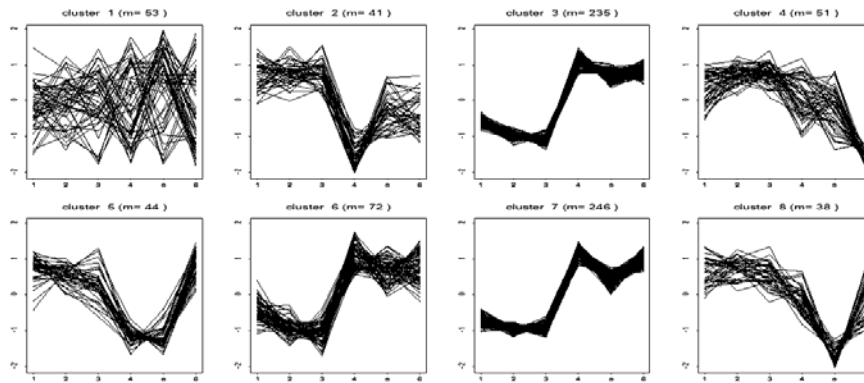


Figure 11. Clustering result of the ensemble averaging method for the LD data with $g = 8$.

Since the hierarchical clustering procedure is used in our ensemble averaging method to cluster the data directly, the resulting clusters are nested. To explore this property, we re-cluster the pre-processed genes into 11 clusters according to the distance matrix produced in the above runs. The results are shown in Figure 12. It is easy to see that cluster 2 in Figure 11 is split into two clusters, clusters 8 and 11 shown in Figure 12; and two new clusters, clusters 5 and 10 shown in Figure 12, are extracted from the noise cluster in Figure 11. This experiment further suggests that if we are interested in the mini-cluster genes, we can make further analysis for the noise cluster genes.

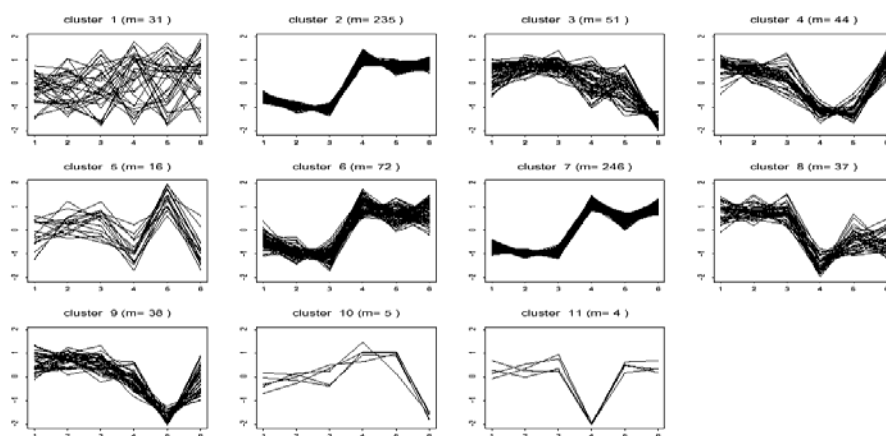


Figure 12. Clustering result of the ensemble averaging method for the LD data with $g = 11$.

5. Discussion

In this paper, we explore the use of the ensemble averaging method for clustering gene expression patterns, and propose a new density estimator for noise genes. The numerical results indicate that the ensemble averaging method can improve the performance of the single model-based clustering method significantly. The ensemble averaging method can automatically separate the scattered, singleton and mini-cluster genes from other genes so that major gene expression patterns can be better identified. In this paper, the ensemble averaging method is demonstrated using four datasets with noise genes. It is obvious that it also works for the datasets without noise genes.

On the Gaussian mixture model-based clustering method, we note that Fraley and Raftery [10] consider various parameterizations of the covariance matrices. For the demonstration purpose, we only consider the most general case where each covariance matrix is an unconstrained positive definite matrix. Extension of our method to other parameterizations is straightforward. In addition, our method can be used with multiple parameterizations; that is, averaging over multiple types of models. Our method also provides a framework for combining the clustering results from different clustering methods, such as K-means, SOM, hierarchical, and others.

On the choice of model range, we suggest to set G_{\min} and G_{\max} according to the BIC values of the models. Let G_{bic} denote the number of clusters of the best BIC model (identified based on multiple runs) for a dataset. In practice, we often set $G_{\min} = G_{bic} - k_1$ and $G_{\max} = G_{bic} + k_2$ with k_1 and k_2 being ranged from 0 to 10. As indicated by Figures 2 and 5 that although BIC is less accurate than the criterion (4) for determining the number of clusters, it can still tell us a rough range where the true number of clusters is. Our experience shows that the ensemble averaging method is not very sensitive to the choices of k_1 and k_2 .

References

- [1] J. D. Banfield and A. F. Raftery, Model-based Gaussian and non-Gaussian clustering, *Biometrics* 49 (1992), 803-821.
- [2] A. Ben-Dor, R. Shamir and Z. Yakhini, Clustering gene expression patterns, *J. Comput. Biol.* 6 (1999), 281-297.
- [3] J. G. Campbell, C. Fraley, F. Murtagh and A. E. Raftery, Linear flaw detection in woven textiles using model-based clustering, *Pattern Recognition Letters* 18 (1997), 1539-1548.
- [4] J. G. Campbell, C. Fraley, D. Stanford, F. Murtagh and A. E. Raftery, Model-based methods for real-time textile fault detection, *Internat. J. Imaging Systems and Technology* 10 (1999), 339-346.
- [5] G. Chen, N. Banerjee, S. A. Jaradat, T. S. Tanaka, M. S. H. Ko and M. Q. Zhang, Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data, *Statist. Sin.* 12 (2002), 241-262.
- [6] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, 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, *Mol. Cell* 2 (1998), 65-73.

- [7] A. Dasgupta and A. E. Raftery, Detecting features in spatial point processes with clutter via model-based clustering, *J. Amer. Statist. Assoc.* 93 (1998), 294-302.
- [8] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. D. Moor and Y. Moreau, Adaptive quality-based clustering of gene expression profiles, *Bioinformatics* 18 (2002), 735-746.
- [9] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc., Ser. B* 39 (1977), 1-38.
- [10] C. Fraley and A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *J. Amer. Statist. Assoc.* 97 (2002), 611-631.
- [11] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach and R. Shamir, An algorithm for clustering cDNA fingerprints, *Genomics* 66 (2000), 249-256.
- [12] S. Hashem, Optimal linear combinations of neural networks, *Neural Networks* 10 (1997), 599-614.
- [13] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein and P. Brown, 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biol.* 1 (2000), 0003.1-0003.21.
- [14] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd Edition, Prentice & Hall International, Inc., 1999.
- [15] L. J. Heyer, S. Kruglyak and S. Yooseph, Exploring expression data: identification and analysis of coexpressed genes, *Genome Res.* 9 (1999), 1106-1115.
- [16] L. Hubert and P. Arabie, Comparing partitions, *J. Classification* (1985), 193-218.
- [17] T. Kohonen, The self-organizing map, *Proc. IEEE* 78 (1990), 1464-1480.
- [18] F. Liang and N. Wang, Dynamic Hierarchical clustering of gene expression profiles, *Pattern Recognition Letters* 28 (2007), 1062-1076.
- [19] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd Edition, Wiley & Sons, 2002.
- [20] G. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, New York, 2000.
- [21] S. Monti, P. Tamayo, J. Mesirov and T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, *Machine Learning* 52 (2003), 91-118.
- [22] M. P. Perrone, Improving regression estimation: averaging methods for variance reduction with extensions to general convex measure optimization, Ph.D. Thesis, Brown University, Rhode Island, 1993.
- [23] W. M. Rand, Objective criteria for the evaluation of clustering methods, *J. Amer. Statist. Assoc.* 66 (1971), 846-850.
- [24] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987), 53-65.

- [25] R. Sharan and R. Shamir, CLICK: a clustering algorithm with applications to gene expression analysis, Proceedings of the 38th Erns Schering Workshop on Bioinformatics and Genome Analysis, H. W. Mewes, H. Seidel and B. Weiss, eds., Springer-Verlag, 2000, pp. 83-108.
- [26] M. Stephens, Dealing with label switching in mixture models, *J. Roy. Statist. Soc., Ser. B* 62 (2000), 795-809.
- [27] 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, *Proc. Natl. Acad. Sci. USA* 96 (1999), 2907-2912.
- [28] G. C. Tseng and W. H. Wong, Tight clustering: a resampling-based approach for identification stable and tight patterns in data, *Biometrics* 61 (2005), 10-16.
- [29] D. H. Wolpert, Stacked generalization, *Neural Networks* 5 (1992), 241-259.
- [30] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo, Model-based clustering and data transformations for gene expression data, *Bioinformatics* 17 (2001), 977-987.
- [31] K. Y. Yeung and W. L. Ruzzo, Principal component analysis for clustering gene expression data, *Bioinformatics* 17 (2001), 763-774.



Kindly return the proof after correction to:

*The Publication Manager
JP Journal of Biostatistics
Pushpa Publishing House
Vijaya Niwas
198, Mumfordganj
Allahabad-211002 (India)*

along with the print charges* by the fastest mail

***Invoice attached**

Proof read by:

Signature:

Date:

Tel:

Fax:

E-mail:

Number of additional reprints required

.....

Cost of a set of 25 copies of additional reprints @ EURO 12.00 per page.

(25 copies of reprints are provided to the corresponding author ex-gratis)