Conference on Applied Statistics in Agriculture 2002 - 14th Annual Conference Proceedings

IMPACT OF DATA TRANSFORMATION ON THE PERFORMANCE OF DIFFERENT CLUSTERING METHODS AND CLUSTER NUMBER DETERMINATION STATISTICS FOR ANALYZING GENE EXPRESSION PROFILE DATA

Guoping Shu

Beiyan Zeng

Deanne Wright

Oscar Smith

See next page for additional authors

Recommended Citation

Shu, Guoping; Zeng, Beiyan; Wright, Deanne; and Smith, Oscar (2002). "IMPACT OF DATA TRANSFORMATION ON THE PERFORMANCE OF DIFFERENT CLUSTERING METHODS AND CLUSTER NUMBER DETERMINATION STATISTICS FOR ANALYZING GENE EXPRESSION PROFILE DATA," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1203

Author Information

Guoping Shu, Beiyan Zeng, Deanne Wright, and Oscar Smith

Impact of Data Transformation on the Performance of Different Clustering Methods and Cluster Number Determination Statistics for Analyzing Gene Expression Profile Data

Guoping Shu, Beiyan Zeng, Deanne Wright, and Oscar Smith

Associative Genetics and Statistical Consulting, Pioneer Hi-Bred Intl, Inc., DuPont Agriculture and Nutrition, 7300 NW 62nd Ave. P.O. Box 1004, Johnston, IA 50131, USA

ABSTRACT

We have assessed the impact of 13 different data transformation methods on the performance of four types of clustering methods (partitioning (K-mean), hierarchical distance (Average Linkage), multivariate normal mixture, and non-parametric kernel density) and four cluster number determination statistics (CNDS) (Pseudo F, Pseudo t^2, Cubic Clustering Criterion (CCC), and Bayesian Information Criterion (BIC)), using both simulated and real gene expression profile data. We found that Square Root, Cubic Root, and Spacing transformations have mostly positive impacts on the performance of the four types of clustering methods whereas Tukey's Bisquare and Interquantile Range have mostly negative impacts. The impacts from other transformation methods are clustering method-specific and data type-specific. The performance of CNDS improves with appropriately transformed data. Multivariate Mixture Clustering and Kernel Density Clustering perform better than K-mean and Average Linkage in grouping both simulated and real gene expression profile data.

Key words: cluster analysis, gene expression profile, data transformation, data normalization, cluster number determination statistics, robustness, Pseudo F, Pseudo t^2, cubic clustering criterion, Bayesian information criterion, Average linkage, k-mean, multivariate mixture-model, kernel density clustering, nonparametric clustering.

1. INTRODUCTION

Clustering analysis plays a pivotal role in grouping and classifying large data sets and identifying co-expressed genes in expression profile data analysis. Because data from expression profile experiments are often large, noisy, asymmetric in distribution, and heterogeneous in scale, data normalization, standardization, and nonlinear transformation are often required in preprocessing data for further statistical analysis. Data transformation is also a widely adopted practice when analyzing a combined data set collected from multiple gene expression experiments or from different profiling technology platforms (Eisen et. al, 1998; Luck et. al, 2001; Yeung et. al, 2001).

The most appropriate data transformation method is not always obvious to experimentalists or data analysts. In this study, we have systematically examined the impact of 13 different data transformation methods on the performance of four types of clustering methods and four cluster number determination statistics using simulated and real gene expression profile data. We have also developed a statistical approach for assessing the performance of different clustering methods for analyzing real expression profile data.

2. METHODS

2.1 Data Transformation Methods

We classify data transformation methods into two types based on their mathematical nature: linear transformation and nonlinear transformation. Linear transformation methods include both normalization and standardization.

Linear Transformation We define linear transformation as any numerical operation that replaces the value of original data y_{ij} with a new value y'_{ij} by adding and/or multiplying by a constant (a , b , $\hat{m}$, $\hat{s}$) through a linear function

$$y'_{ij} = a\frac{(y_{ij}-\hat{\mu}_i)}{\hat{\sigma}_i}+b \qquad (1)$$

By the above definition, data normalization and data standardization are two different types of linear transformations with the former operating on either $\hat{m}$ or $\hat{s}$ and the latter on both $\hat{m}$ and $\hat{s}$. The types of linear transformation included in this study are listed in Table 1.

The five data standardization methods we discuss (Table 1) include procedures that are based on either L-estimators (Standardizing (STD), Interquartile Ranging (IQR)), M-estimators (Tukey's bisquare (TBS) and Huber transformation), and density estimate (Spacing) of location and scale (Table 1, Tukey et al., 1977; Hoaglin et al., 1983; Wilcox, 1997). Given a sample $X_1, X_2, \mathrm{L}\ , X_n$ from a population with a true standard deviation of s , the L-estimators (STD, IQR) minimize the general functions $f_1(\hat{\mu}) = \sum_{i=1}^{n}(\frac{X_i-\hat{\mu}}{\sigma})^2$ and $f_2(\hat{\mu}) = \sum_{i=1}^{n}\left|\frac{X_i-\hat{\mu}}{\sigma}\right|$ respectively. The M-estimators (TBS and Huber) minimize the general function $f_3(\hat{\mu}) = \sum_{i=1}^{n}\Psi\left(\frac{X_i-\hat{\mu}}{\sigma}\right)$, where Y is a weight function and the maximum likelihood estimate of $\hat{m}$ is the robust measure of location. Among various weight functions available in literature, we choose the two that are most widely used: Tukey's bisquare function ($\Psi_T(x)$) and Huber's function ($\Psi_H(x)$). These functions are defined as:

$$\Psi_T(x) = \begin{cases} x(c^2-x^2)^2 & |x| \le c \\ 0 & |x| > c \end{cases} \quad \text{and} \quad \Psi_H(x) = \begin{cases} x & |x| < c \\ \mathrm{sign}(x)c & |x| \ge c \end{cases}$$

where sign(x) is equal to -1, 0, or 1 depending on the sign of x and c is a tuning constant. We use median absolute deviation (MAD) to approximate s in calculation of Ψ functions as suggested by Hoaglin et al. (1983).

The two robust measures of scale (s), Tukey's bisquare A estimate and Huber t estimate (see Table 1) are defined as

$$A = kS_M \frac{\sqrt{n\sum_i \Psi^2(Y_i)}}{\left|\sum_i \Psi'(Y_i)\right|} \quad \text{and} \quad \tau = kS_M \sqrt{\tfrac{1}{n}\sum_i \rho(Y_i)} \qquad (2)$$

where S_M is a scale estimate from a sample of size n, k is a constant, and r is a weight function (Table 1; Tukey et al., 1977; Hoaglin et al., 1983; Wilcox, 1997). The Spacing transformation (Table 1) is based on a density smoothing technique (Jannsen, 1995; Shu et al., submitted).

Nonlinear Transformation Among numerous types of the nonlinear transformation procedures available in literature, we consider four of the most commonly used ones for gene expression profile analysis (see Table 1). These four all belong to the family of power transformations given by

$$y'_{ij} = \begin{cases} ay_{ij}^{p} + b & (p \neq 0) \\ c\log_m y_{ij} & (p = 0) \end{cases} \qquad (3)$$

In both function (1) and (3), a, b, c, d, and p are real numbers and we require $a > 0$ for $p > 0$ and $a < 0$ for $p < 0$ (Hoaglin et al., 1983).

Table 1. Linear and Nonlinear Transformation Methods

Types of Transformation	Methods	Location ($\hat{m}$)	Scale ($\hat{s}$)
Linear: Normalization	Mean Centering Median Centering Norm Weighting (Norm) SD Weighting (USTD)	mean median 0 0	1 1 vector length standard deviation
Linear: Standardization	Standardizing (STD) Interquartile Rang (IQR) Tukey's Bisquare (TBS) Huber Transformation Spacing	mean median TBS M estimate Huber M estimate mid minimum-spacing	standard deviation interquartile range TBS A estimate Huber t estimate minimum spacing
*Nonlinear: Power Transformation	$\log_2$ $\log_{10}$ Squared Root (SQRT) Cubic Root (CURT)	($m = 2$) ($m = 10$) ($p = \frac{1}{2}$) ($p = \frac{1}{3}$)	

* The location and scale are not required to be specified thus are not listed

2.2 Clustering Methods

Various types of clustering methods have been applied to gene expression profile data analysis, such as, hierarchical clustering (Eisen *et al.*, 1998), self-organizing maps (Tamayo *et al.*, 1999), k-means (Tavazoie *et al.*, 1999), multivariate mixture model-based methods (Fraley and Raftery, 1998; Yeung *et al.*, 2001; Ghosh and Chinnaiyan, 2002), and non-parametric kernel density clustering (Shu et al., submitted). Included next are brief descriptions of the four types of clustering methods whose performance we assess.

2.2.1 Kernel Density Clustering

Kernel density clustering (Shu et al., submitted) uses kernel smoothing techniques to determine the modes or local maxima of the density function given by

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{4}$$

Where n is the total number of observations in the data set, K is a kernel (or weight) function, h is the bin width, or smoothing parameter for any observation x in a bin that centered at an observation X_i. The distance or dissimilarity measure between two objects (or clusters), x_i and x_j is computed by

$$d(x_i, x_j) = \begin{cases} \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) & \text{if } d(x_i, x_j) \le r \\ \infty & \text{otherwise} \end{cases} \tag{5}$$

where r is the radius of a closed hypersphere centered at point x, $f(x)$ is the estimated density at x (Silverman, 1986; Scott, 1992; SAS, 1999).

2.2.2 Mixture Model-based Clustering

Mixture model-based clustering is based on the theory of finite mixture distribution. A finite mixture distribution is a linear function of a number of component probability distributions. When every component distribution for each group or true cluster k is a multivariate normal distribution, the finite mixture distribution is called Multivariate Normal Mixture (MNM),

$$\begin{aligned} f(X_i; p_k, \mu_k, \Sigma_k) &= p_1 f_1(X_i) + p_2 f_2(X_i) +, \mathrm{L} \;, + p_G f_G(X_i) \\ &= \sum_{k=1}^{G} p_k f_k(X_i; \mu_k, \Sigma_k) \end{aligned} \tag{6}$$

where G is the number of groups or true clusters in the population, p_k is the proportion of group k in the population or the probability that an object x_i belongs to population k, and $f_k(X_i; \mu_k, \Sigma_k)$ denotes the multivariate normal density function of component k. More specifically, for a data set (a N x M coordinate data matrix (A_{ij})) that contains G groups ($k = 1, 2, ..., G$), the probability density for object i in group k is

$$f_k(X_i; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp^{-(X_{ik} - \mu_k)'\Sigma_k^{-1}(X_{ik} - \mu_k)/2} \tag{7}$$

where Σ_k is the M x M within-group variance-covariance matrix of group k, $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is its inverse matrix. Equation (7) is also called the kth component multivariate normal density function.

From (6) and (7), the likelihood of which N objects are sampled from k groups given parameter θ can be described using the likelihood function

$$L(\theta; X) = \prod_{i=1}^{N} f(X_i; p_k, \mu_k, \Sigma_k) \tag{8}$$

and the log likelihood function is

$$\begin{aligned} l(\theta) = \log L(\theta; X) &= \sum_{i=1}^{N} \log\left[\sum_{k=1}^{G} p_k f_k(X_i; \mu_k, \Sigma_k)\right] \\ &= \sum_{i=1}^{N} \log\left[\sum_{k=1}^{G} p_k \frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp^{-(X_{ik}-\mu_k)'\Sigma_k^{-1}(X_{ik}-\mu_k)/2}\right] \end{aligned} \tag{9}$$

In this study, we use maximum likelihood to identify the unknown mixture component origin for object x_i and estimate the population parameters $\theta = (p_1, \mathrm{L}\ , p_G;\ \mu_1, \mathrm{L}\ , \mu_G;\ \Sigma_1, \mathrm{L}\ , \Sigma_G)$ using EM algorithms proposed by Fraley (1998) and implemented in a R language software package Mclust (Fraley and Raftery, 1999).

We have assessed six different mixture models summarized in Banfield and Raftery (1993) and Fraley (1998). The six models are Sum of Squares Model (Trace, EI), Unconstrained Σ_k Model (Unconstrained, VVV), Minimum Determinant Model (Determinant, EEE), Spherical Cluster Model (Spherical, VI), Murtagh-Raftery Model (S, EEV), and Banfield-Raftery Model (S*, VEV) (Banfield and Raftery, 1993; Fraley and Raftery, 1998). Each of the above models imposes different constraints to the within-group sample variance-covariance matrix Σ_k of (9). The modeling and data analysis were done using Mclust (Fraley and Raftery, 1999).

2.2.3 Average Linkage and K-mean Methods Both Average linkage and K-mean clustering methods have been widely used for years and are well documented. We used Euclidean distance in both Average linkage and K-mean clustering and an adaptive updating algorithm for K-mean (Gordon, 1999, SAS, 1999).

2.3 Cluster Number Determination Statistics (CNDS)

Several statistical criteria for detecting or determining the number of groups or true clusters in a data set exist in the literature (Gordon, 1999; Milligan and Cooper, 1985).

We assess the impact of data transformation on the performance of four cluster number determination statistics.

(1) Pseudo F : a statistic first proposed by Calinski and Harabasz (1974) and thus also called Calinski-Harabasz test, which is defined as

$$Pseudo\ F = \frac{\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2 - \sum_{k}^{G}\sum_{i=1}^{n_k}(X_i - \overline{X}_k)^2}{G-1}}{\dfrac{\sum_{k}^{G}\sum_{i=1}^{n_k}(X_i - \overline{X}_k)^2}{n-G}} \tag{10}$$

where n_k are number of objects in cluster $k\,(1 < k < n)$, $n - G = \sum_{k=1}^{G}(n_k - 1)$, and the x_i and $\bar{x}_k$ are 1 x m observation vector for object i and the centroid (the mean vector) for group k respectively at any level of joining.

(2) Pseudo t^2: a statistic for testing whether or not joining two clusters (A and B) into a new cluster (U) is statistically meaningful by checking the change in the sum of squares,

$$Pseudo\ t^2 = \frac{SS_{AB}}{\frac{SS_A + SS_B}{n_A + n_B - 2}} = \frac{SS_U - SS_A - SS_B}{\frac{SS_A + SS_B}{n_A + n_B - 2}} \qquad (11)$$

This test was originally proposed in a different form by Duda and Hart (1973) and was called Je(2)/Je(1) test.

(3) CCC Test: a statistic based on the assumption that a uniform distribution on a hyperrectangle will be divided into clusters shaped roughly like hypercubes. CCC is based on simulation result under the null hypothesis of multivariate uniform distribution. It can be viewed as the product of two items

$$CCC = \ln\left[\frac{1 - E(R^2)}{1 - R^2}\right] \cdot \frac{\sqrt{(np/2}}{(0.001 + E(R^2))^{1.2}} \qquad (12)$$

where R^2 represents the proportion of variance explained by clusters, $E(R^2)$ is its expected value under the null hypothesis, and p is an estimate of the dimensionality of between cluster variation and n is the total number of objects. Consult Sarle (1983) for details about the interpretation of the CCC.

(4) Bayesian Information Criteria (BIC): A statistic used for determining the number of clusters detected by mixture model clustering. We use the formula from Schwarz (1978):

$$BIC_k = 2\log p(D/M_k) \text{ » } 2\log p(D/\hat{q}_k, M_k) - v_k \log(n) \qquad (13)$$

where v_k is the number of parameters to be estimated in a model, M_k and $\hat{q}_k$ are the maximum likelihood estimates for the parameter vectors in the model.

2.4 Measuring Between-cluster Separation and Within-cluster Coherence

We used accumulated between-cluster R^2 to measure the between-cluster separation or isolation, which is defined as

$$R^2 = 1 - \frac{\text{Total within-cluster sum of squares}}{\text{Total sum of squares}} = 1 - \frac{\sum_{k=1}^{G}\sum_{i=1}^{n_k}(X_i - \bar{X}_k)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad (14)$$

where G is the number of clusters determined by CNDS and n_k is the number of objects (genes) in cluster k.

We use profile plots to examine the within-cluster coherence. Figure 6 and Figure 7 are illustrations of profile plots. The X-axis could be different time-points, developmental stages, or different treatments etc., and Y-axis is level of gene expression such as signal density.

2.5 Measuring Performance of a Clustering Method

A widely adopted method for assessing the performance of a clustering method is external validation (Gordon, 1999). The procedure of external validation we employ includes the following steps: (1) a data set that has K known clusters is generated using Monte Carlo simulation, each observation (gene) vector in the data set carries a cluster membership ID, called design cluster ID, this data set is used as an external standard for comparison; (2) the same data set is partitioned or grouped into K clusters by the clustering method to be assessed and each observation in the data set is assigned a new cluster ID, called assigned cluster ID; (3) The degree of match or resemblance between the assigned and the designed cluster ID is computed using a match coefficient, called Hubert-Arabie Adjusted Rand Index (Hubert and Arabie, 1985; Rand, 1971), which is given as

$$R_{HA} = \frac{\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \sum_{i=1}^{c_1}\binom{n_{i.}}{2}\sum_{j=1}^{c_2}\binom{n_{.j}}{2}/\binom{n}{2}}{\left[\sum_{i=1}^{c_1}\binom{n_{i.}}{2}+\sum_{j=1}^{c_2}\binom{n_{.j}}{2}\right]/2-\sum_{i=1}^{c_1}\binom{n_{i.}}{2}\sum_{j=1}^{c_2}\binom{n_{.j}}{2}/\binom{n}{2}} \qquad (15)$$

where $n_{i.} = \sum_{j=1}^{c_2} n_{ij}$ and $n_{.j} = \sum_{i=1}^{c_1} n_{ij}$, and c_1, c_2 are the number of clusters in the two partitions respectively. $R_{HA} = 1$ indicates a perfect match, $R_{HA} = 0$ indicates a random grouping or a complete failure of the clustering method in recovering the known or designed clusters. Thus the Adjusted Rand Index measures the rate of cluster identity recovery of a clustering method.

2.6 Measuring the Impact of Data Transformation on Clustering Method and CNDS

We used the following equation to measure the impact of a data transformation method on the performance of a clustering method (IM):

$$IM = R_{HA,t} - R_{HA,o} \qquad (16)$$

where $R_{HA,t}$ is the adjusted Rand Index computed from the transformed data using equation (15), and $R_{HA,o}$ is the Adjusted Rand Index from the original or untransformed data (the first column in both Table 1 and Table 2). A negative IM value ($IM < 0$) indicates that the data transformation reduce the rate of cluster identity recovery of the clustering method and a positive IM value ($IM > 0$) indicates that the transformation improves the rate of cluster identity recovery.

The impact of data transformation method on the performance of a cluster number determination statistics (CNDS) was measured by peak shift in a profile plot such as the one shown in Figure 1A. The cluster number on the X-axis that corresponds to the peak is an assigned number of clusters by the CNDS. Because the number of clusters in the simulated data is known (11 clusters for data set A and 15 clusters for data set B), we will say that a data transformation method has a positive impact on a CNDS if the CNDS can detect the known cluster number from the transformed data, and a data transformation method is said to have a negative impact if it can detect the known cluster number from the original data but fail to do so from the transformed data.

3. DATA SETS

3.1 Simulated Data

We generated two expression profile data sets using Monte Carlo simulation. The key specifications for each data set are:

Data Set A: 750 genes (objects), each has10 observations (variables) representing 10 sampling time points of a developmental progression. There are 11 clusters with cluster size range from 20 to 150 genes per cluster. Each cluster has a unique linearly or nonlinearly increasing or decreasing trend or curve from time point 1 to time point 10.

Data Set B: there are 1040 genes, each has 10 observations (variables) representing 10 sampling time points of a developmental progression, 15 clusters with cluster size ranging from 20 to 150 genes per cluster. Each cluster simulates a cyclic or non-cyclic (time point-specific or development stage-specific) expression profile. That is, the level of gene expression either oscillates across development time points or only goes up or down at one or two time points and stays stationary at other stages.

3.2 Real Expression Profile Data

We used an expression profile data set from Lee et al., (2002). The data set we used for this analysis has 1130 genes and the level of gene expression was measured using microarray at 5 developmental stages of kernel development (5, 10, 15, 20, 25 days after pollination (DAP)). See Lee et al. (2002) for detail.

4. RESULTS AND DISCUSSION

4.1 Performance of Different Clustering Methods in Analyzing Untransformed Data

We first assessed the performance of four types of clustering methods by applying each method to two simulated data sets that have not undergone any data transformation. The Adjusted Rand Index, which measures the rate of cluster identity recovery of a clustering method, is reported in the first column labeled as "Original" in both Table 2 and Table 3. The results show that the three mixture clustering methods and the kernel density method perform the best in clustering both data sets. The average linkage method performs very well in clustering data set B but very poorly in clustering data set A. K-mean method does not perform well in clustering either data set.

Table 2. Adjusted Rand Index for Untransformed and Transformed Data Set A By Different Clustering Methods

Normalizatiion	Original	Mean	Median	Norm	USTD
Density	1.00	1.00	1.00	0.89	0.89
Mixture EI	0.96	0.96	0.96	0.69	0.69
Mixture VI	1.00	1.00	1.00	0.96	0.79
Mixture VEV	1.00	1.00	1.00	0.99	0.99
K-Mean	0.79	0.79	0.79	0.69	0.69
Average Linkage	0.43	0.43	0.43	0.19	0.19

Standardization	STD	IQR	TBS	Huber	Spacing
Density	0.88	0.84	1.00	1.00	1.00
Mixture EI	0.69	0.72	0.83	0.73	0.96
Mixture VI	0.86	0.87	1.00	0.99	1.00
Mixture VEV	0.99	1.00	1.00	1.00	1.00
K-mean	0.59	0.72	0.76	0.71	0.86
Average Linkage	0.19	0.32	0.38	0.38	0.49

Transformation	Log_2	Log_{10}	SQRT	CURT
Density	0.87	0.87	1.00	1.00
Mixture EI	0.68	0.68	0.95	0.89
Mixture VI	0.84	0.86	1.00	1.00
Mixture VEV	0.77	0.85	1.00	1.00
K-mean	0.70	0.70	0.79	0.79
Average Linkage	0.29	0.29	0.61	0.60

4.2 Impact of Data Transformation on Performance of Different Clustering Methods

The Adjusted Rand Index, which measures the resemblance between cluster ID obtained from the transformed data and the designed ID in the original data (the first column in both Table 1 and Table 2), are reported in the columns after the "Original" column in Table 2 and Table 3. The impact of each data transformation method was measured using function (16) described in Section 2.6. There are several patterns obvious to both data set A and B: (1) Mean-centering and Median-centering have no impact on the performance of any clustering method (IM=0) (2) Spacing, SQRT, and CURT transformation have mostly positive impact ($IM \geq 0$) whereas TBS and IQR have a mostly negative impact ($IM \leq 0$) and (3) $\log_2$ and $\log_{10}$ transformation have the same impact (identical Adjusted Rand Index) on all clustering methods except Mixture VI and Mixture VEV. The impact of other 9 transformation methods 9 is clustering method-specific and data type-specific. For instance, the $\log_2$, $\log_{10}$, and STD transformation have a mostly negative impact in data set A but have either no impact or a positive impact in data set B except for Average Linkage clustering where both $\log_2$ and $\log_{10}$ transformation have large negative impact on the performance of average linkage clustering (IM =0.37-0.99= -0.62).

4.3 Impact of Data Transformation on the Performance of Clustering Number Determination Statistics (CNDS)

We have assessed the impact of all 13 data transformation methods on the performance of different cluster number determination statistics (Section 2.3). Due to the limit in space, here we only show the results from four commonly used data transformation methods: USTD, STD, TBS, and $\log_2$ applied to data set B. Because different CNDS are formulated on different statistical models, and might not be suitable to measuring the performance of every clustering method we are studying, we only show the results from the most appropriate CNDS-Clustering Method combinations. They are pseudo F, CCC, and Pseudo t^2 for average linkage, pseudo F and CCC for K-mean, Pseudo t^2 and BIC for Mixture model-based clustering. The results are shown in Figures 1 to 4.

The ways we measured the impact of data transformation were described in Section 2.6.

Table 3. Adjusted Rand Index for Untransformed and Transformed Data Set B By Different Clustering Methods

Normalizatiion	Original	Mean	Median	Norm	USTD
Density	0.93	0.93	0.93	0.97	0.99
Mixture EI	0.99	0.99	0.99	0.99	0.99
Mixture VI	0.99	0.99	0.99	0.99	0.99
Mixture VEV	0.99	0.99	0.99	0.99	0.99
K-Mean	0.64	0.64	0.64	0.64	0.64
Average Linkage	0.99	0.99	0.99	1.00	1.00

Standardization	STD	IQR	TBS	Huber	Spacing
Density	0.97	0.89	0.89	0.96	0.97
Mixture EI	0.99	0.99	0.98	0.99	0.99
Mixture VI	0.99	0.99	0.99	0.99	0.99
Mixture VEV	0.99	0.99	0.99	0.99	0.99
K-mean	0.86	0.59	0.59	0.85	0.59
Average Linkage	0.99	0.99	0.99	1.00	1.00

Power Transformation	LOG_2	LOG_{10}	SQRT	CURT
Density	0.97	0.97	1.00	0.93
Mixture EI	0.99	0.99	0.99	0.99
Mixture VI	0.99	0.94	0.99	0.99
Mixture VEV	0.99	0.94	0.99	0.99
K-mean	0.81	0.81	0.73	0.76
Average Linkage	0.37	0.37	1.00	0.99

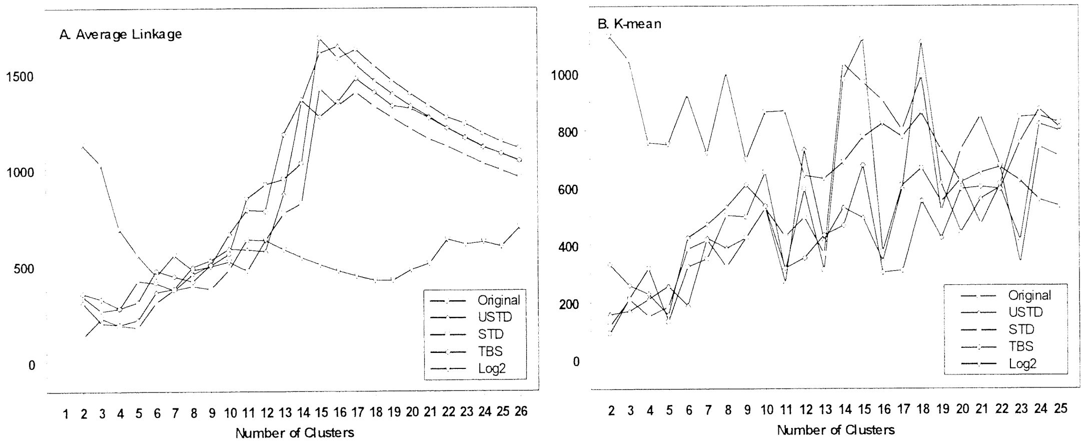
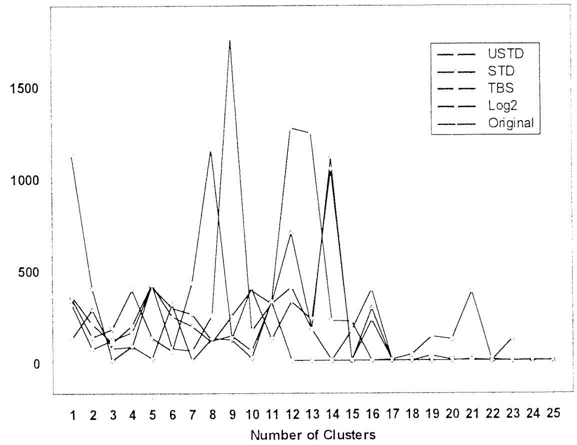


Figure 1. Impact of Data Transformation on the Performance of Pseudo F Statistics (Y-axis) When Applied to Average Linkage (A) and K-mean (B) Clustering

It should be noted that the different ways of finding a cluster number cutoff point which indicates the number of clusters in the data on CNDS profile graphs. On Pseudo F, CCC, and BIC graph, it is the number on the X axis that correspond to a peak reading on Y axis, which we define as k; on the Pseudo t^2 graph, the cutoff point is $k+1$ instead of k. Pseudo t^2 measures the relative degree of increase of within-cluster sum of squares (or within-cluster heterogeneity) at joining to form k clusters from $k+1$ clusters in hierarchical clustering. A peak Pseudo t^2 value (Y-axis reading) at k (X-axis reading) together with a valley (low reading) at $k+1$ suggests that the joining into k clusters from $k+1$ clusters creates a highly heterogeneous new cluster and thus is not desirable. Therefore, the correct cutoff point should be $k+1$. For instance, in Figure 2, Pseudo t^2 value peaks at cluster 14 and drops at cluster 15 for both original and STD transformed data, thus we say that the correct cutoff point is 15 or say that the Pseudo t^2 reading suggests the existence of 15 clusters in the data. Similarly, the Pseudo t^2 suggests 10 clusters for TBS transformed data and 14 clusters for USTD transformed data.

Figure 2. Impact of Data Transformation on the Performance of Pseudo t^2 Statistics (Y-axis) When Applied to Average Linkage Clustering



The results shown in the four figures (Figure 1- Figure 4) can be summarized as the follows: (1) for average linkage clustering of the original and STD transformed data (Figure 1A, Figure 2, and Figure 3A), pseudo F, Pseudo t^2, and CCC all detect 15 clusters, the correct number by design for data set B (2) for K-mean clustering of the original data and TBS transformed data, both pseudo F and CCC detect 15 clusters (Figure 1B, 3B). (3) for multivariate normal mixture clustering using EVE model, BIC fails to detect the correct cluster number from the original data (Figure 4A), but detect the correct number from USTD and STD transformed data (Figure 4B). For better visualization, we use relative BIC (Relative$\text{BIC}_i = \text{BIC}_i / |\min \text{BIC}_i|$) in Figure 4B. In Figure 4B, we also see that the two curves for USTD and STD are completely overlapping to each other because their BIC values are the same.

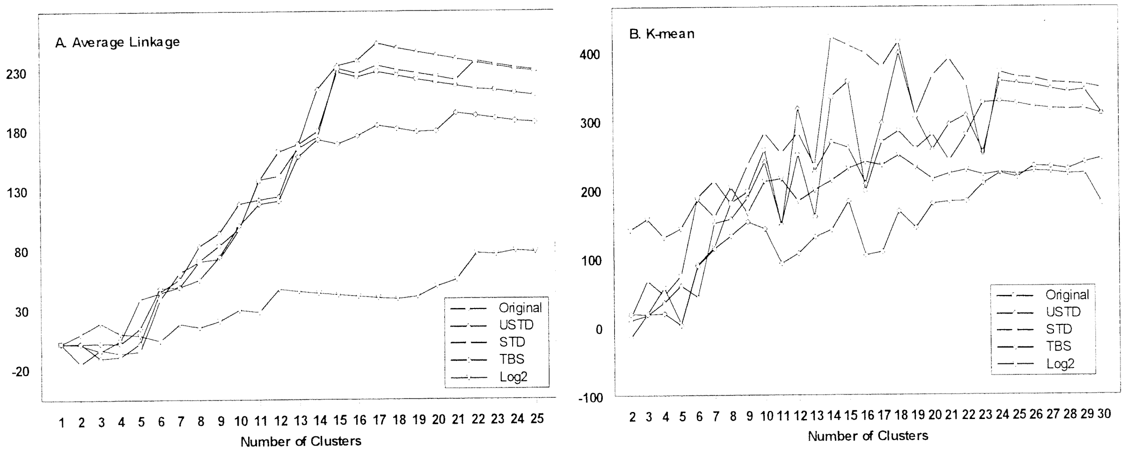


Figure 3. Impact of Data Transformation on the Performance of Cubic Clustering Criterion Statistics (CCC, Y-axis) When Applied to Average Linkage (A) and K-mean (B) Clustering

For the six mixture clustering models described in section 2.2, we also assess the impact of model selection on the performance of CNDS. Figure 4A shows the BICs of mixture clustering of the original data using six different statistical models. The BICs peak at cluster 15 for EI, VI, and EEE models, but not for VEV model, which peak at 8 clusters, and other two models (EEV, and VEV). However, after transforming the data with USTD and STD, the BIC peak shifts to cluster 15. The above results show that appropriate data transformation does improve the performance of CNDS.

We do not find an association between the impact on rate of cluster identity recovery and the impact on CNDS for any data transformation method. A data transformation method could have negative impact on the rate of cluster identity recovery but have no impact or have positive impact on CNDS. For instance, TBS transformation has a strong negative impact on the rate of clustering identity recovery in K-mean clustering (IM = 0.59-0.86 = -0.27) (Table 3), but it is the only data transformation method by which both Pseudo F and CCC identify the correct cluster number in K-mean clustering (Figure 1B, 3B).

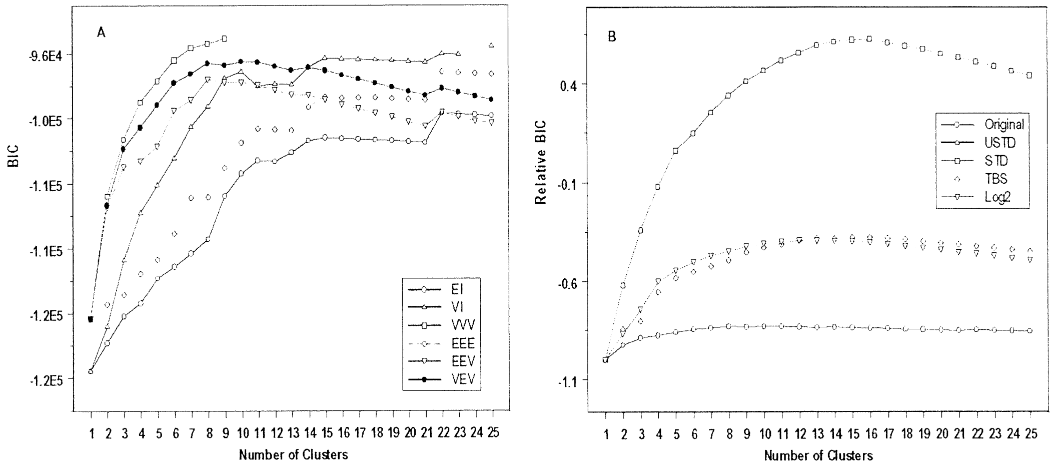


Figure 4. Impact of Data Transformation on the Performance of Bayesian Information Criterion (BIC) When Applied to Mixture Model Clustering (A) Original (Untransformed) Data Were Clustered Using Six Different Mixture Model-based Clustering Procedures (B) Original and Four Transformed Data Were Clustered by Mixture Model-based Clustering Procedure (VEV Model)

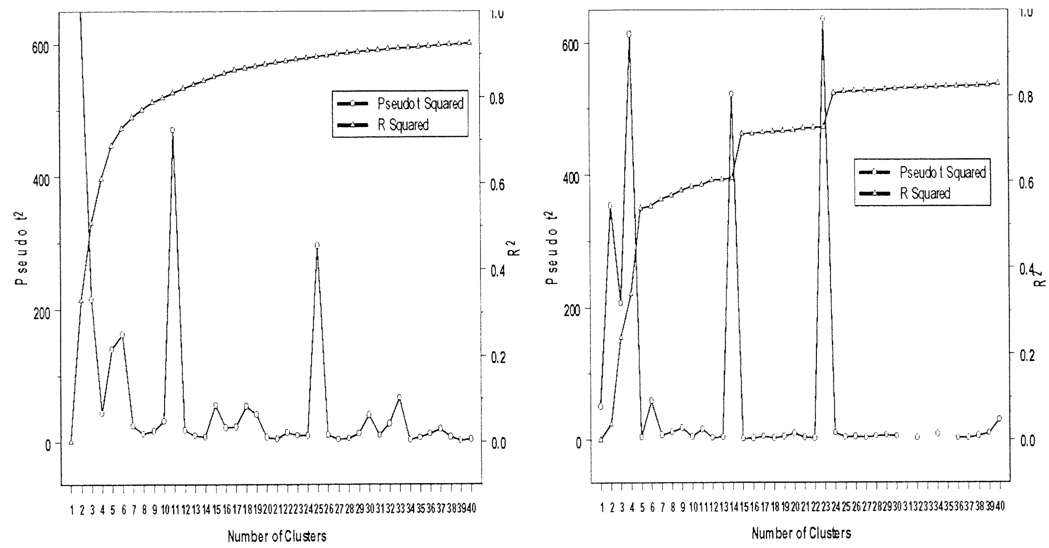


Figure 5. The number of clusters in maize embryo expression profile data determined with Pseudo t^2 and R^2 for (A) Multivariate Mixture Clustering (EI model) and (B) Average Linkage clustering

4.4 Assessing the Performance of Different Clustering Methods in Analyzing Real Expression Profile Data

We have compared the performance of the above four types of clustering methods in analyzing several real expression profile data sets. Here we only report the comparison between Average Linkage clustering and Multivariate Normal Mixture clustering (EI model) in analyzing a microarray data from maize embryo development of Lee et al. (2002). See Section 3 for detail information about the data set.

An intrinsic difficulty in assessing the performance of different clustering methods based on results of analyzing real data is that the number of true clusters and cluster identity of each object (gene) in real data are unknown and the Adjusted Rand Index can not be computed and the assessment methodology we used for simulated data will not apply. Here we propose an different approach that measures the degree of between-cluster separation (isolation) using CNDS and accumulated between-cluster R^2 and measure the within-cluster coherence (agreement in profile pattern) using profile plot. We analyzed the maize data set using this approach and the results are shown in Figure 5, and 6, and 7.

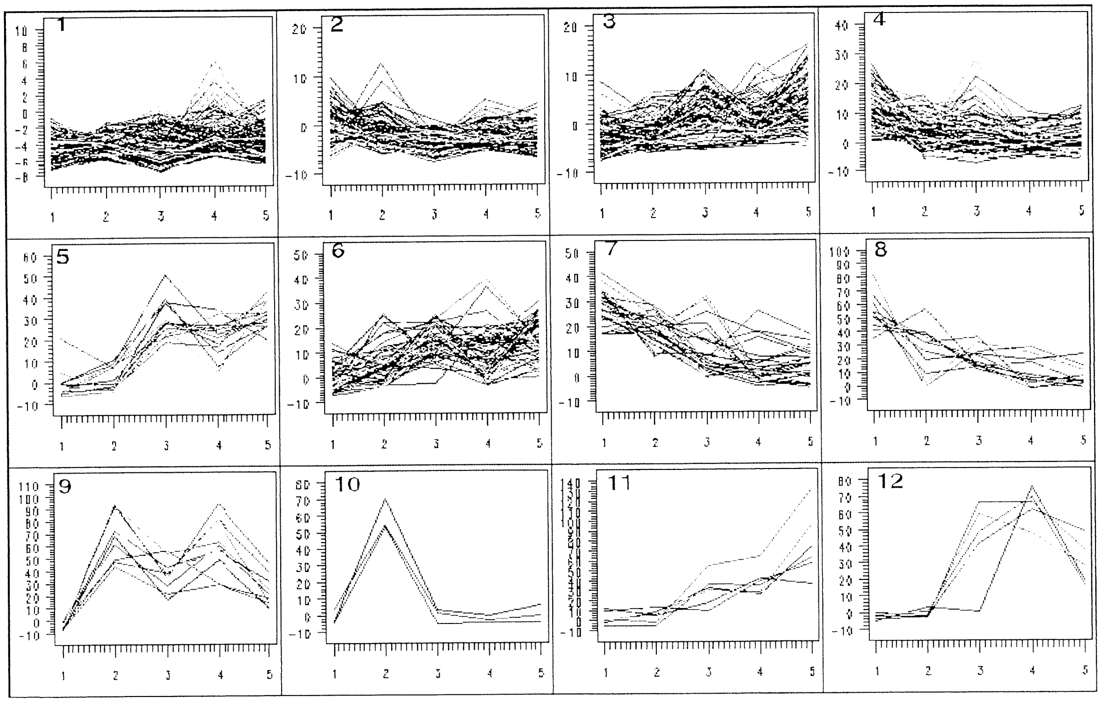


Figure 6. The Expression Profile of 12 clusters of Maize Genes Identified Using Mixture EI Method

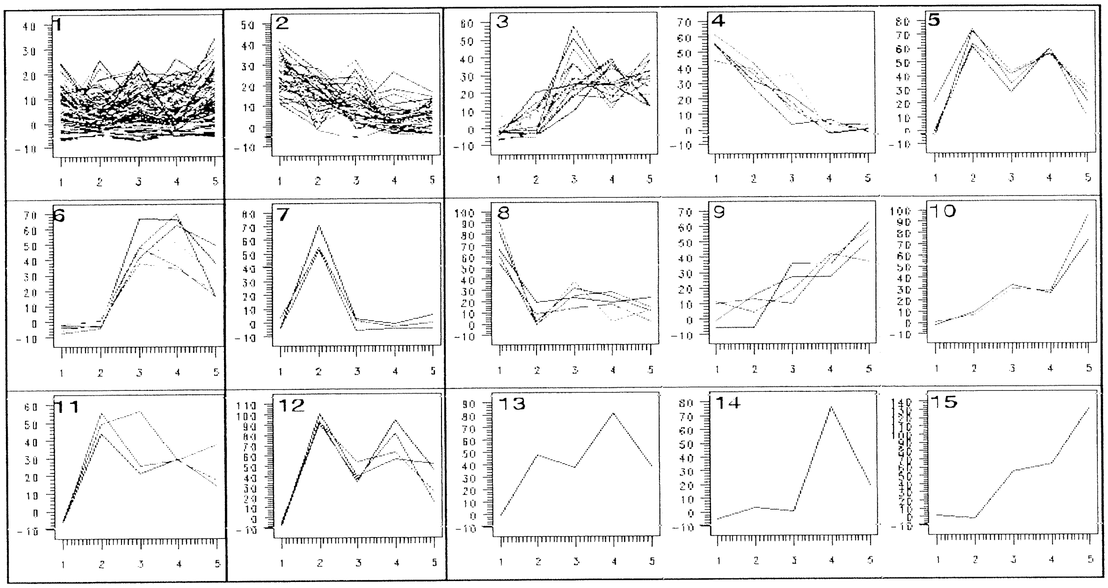


Figure 7. Expression Profiles of 15 clusters of Maize Genes Identified Using Average Linkage Method

The possible cutoff points in the maize ear data are 4, 7, 12 and 26 clusters for Mixture EI clustering and 5, 15 and 24 clusters for Average Linkage clustering (Figure 5). The corresponding R^2 values are 0.6114, 0.7531, 0.8209, and 0.8975 respectively for Mixture EI, and 0.5380, 0.7110, and 0.8051 respectively for Average-linkage (Figure 5). The R^2 values indicate that the clusters delineated by Mixture EI clustering have better between-cluster separation. The profiles generated from 12-cluster cutoff for Mixture EI method and that from 15-cluster cutoff for Average Linkage are shown in Figure 6 and Figure 7, respectively. From the two figures we can see that cluster 1 generated by Average Linkage clustering (Figure 7) is highly heterogeneous and is separated into 5 clusters of different trends of expression by Mixture EI clustering (cluster 1, 2, 3, 4, 6 in Figure 6). We can also see that Cluster 11 (Figure 6) generated by Mixture EI is coherent but is erroneously broken down into three clusters by Average Linkage clustering (cluster 9, 10, 15 in Figure 7).

5. SUMMARY

Data transformation has a measurable impact on the performance of both clustering methods and clustering number determination statistics (CNDS). Square Root, Cubic Root, and Spacing transformations have mostly the positive impact whereas Tukey's Bisqure and Interquantile Range have mostly negative impacts. The impacts from other transformation methods are clustering method-specific and data type-specific. The performance of CNDS improves with appropriately transformed data. Multivariate Mixture Clustering and Kernel Density Clustering perform better than K-mean and Average Linkage in grouping both simulated and real data.

ACKNOWLEDGEMENTNS

We thank Antoni J. Rafalski for providing the maize microarray data and suggestions and comments on data analysis. We also thank Mark Cooper, Stan Luck, and Kevin Wright for many helpful discussions.

REFERENCES

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803–821.
Calinski, T. and Harabasz J. (1974) A dendrite method for cluster analysis. Communications in Statistics, 3, 1–27.
Duda, R.O. and Hart, P.E. (1973) Pattern Classification and Scene Analysis, New York: John Wiley & Sons, Inc.
Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Science USA, 95, 14863–14868.
Fraley C. (1998) Algorithms for model-based Gaussian hierarchical clustering. SIAM Journal on Scientific Computing, 20, 270-281.
Fraley, C. and Raftery, A. E. (1998) How many clusters? Which clustering method? -Answers via model-based cluster analysis. The Computer Journal, 41, 578–588.
Fraley, C. and Raftery, A. E. (1999) Mclust: Software for model-based cluster analysis. Journal of Classification, 16, 297-306.
Ghosh D. and Chinnaiyan A. M. (2002) Mixture modeling of gene expression data from microarray experiments. Bioinformatics, 18, 275-286.
Gordon A. D. (1999) Classification, 2nd edition, Chapman & Hall/CRC.
Hastie T., Tibshirani R., and Friedman J. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.
Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983) Understanding Robust and Exploratory Data Analysis. New York: John Wiley & Sons, Inc.
Hubert, L. and Arabie, P. (1985) Comparing partitions. Journal of Classification, 2, 193–218.

Jannsen P., Marron J. S., veraverbeke, N, and Sarle, W. S. (1995) Scale measures for bandwidth selection. J. of Nonparametric Statistics, 5, 359-380.

Lee J., Williams M. E., Tingey S. V. and Rafalski J. A. (2002) DNA array profiling of gene expression changes during maize embryo development. Funct. Integr. Genomics, 2, 13-27.

Li L., Weinberg C. R., Darden T. A., and Pedersen L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 17, 1131-1142.

Luck S. (2001) Normalization and error estimation for expression profiles. In Microarrays: Optical Technologies and Informatics, M. L. Bittner, Y. Chen, A. N. Dorsel, Edward R. D. Edited, Proceedings of SPIE Vol. 4266, 153-157.

Milligan, G.W. and Cooper, M.C. (1985) An examination of procedures for deter-mining the number of clusters in a data set. Psychometrika, 50, 159–179.

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

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

SAS Institute Inc., SAS/STAT User's Guide, Version 8, Cary, NC: SAS Institute Inc., 1999.

Sarle, W.S. (1983) Cubic Clustering Criterion, SAS Technical Report A-108, Cary, NC: SAS Institute Inc.

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

Scott, D.W. (1992) Multivariate Density Estimation: Theory, Practice, and Visualization, New York: John Wiley & Sons, Inc.

Shu G., Zeng B., Chen P., and O. Smith (2002) Non-parametric Kernel Density Clustering for gene expression data. (submitted).

Silverman, B.W. (1986) Density Estimation, New York: Chapman and Hall.

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

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

Tukey, J.W. (1977). Exploratory Data Analysis. Reading, Massachusetts: Addison Wesley Publishing Company.

Wilcox, R. R. (1997). Introduction to Robust Estimation and Hypothesis Testing. San Diego: Academic Press.

Yeung, K. Y., Fraley C., Murua A., Raftery A. E. and Ruzzo, W. L. (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics, 17, 977-987