A New Approach to Cluster Analysis: the Clustering-function Based Method

Baibing Li
Business School, Loughborough University
Loughborough, LE11 3TU, UK
email: b.li2@lboro.ac.uk

Abstract

The purpose of this paper is to present a new statistical approach to hierarchical cluster analysis with n objects measured on p variables. Motivated by the model of multivariate analysis of variance and the method of maximum likelihood, a clustering problem is formulated as a least squares optimisation problem, simultaneously solving for both an n-vector of *unknown* group membership of objects and a linear clustering function. This formulation is shown to be linked to linear regression analysis and Fisher linear discriminant analysis and includes principal-component regression for tackling multicollinearity or rank-deficiency, polynomial or B-splines regression for handling non-linearity, and various variable selection methods to eliminate redundant variables from data analysis. Algorithmic issues are investigated using sign eigenanalysis.

Keywords: Discriminant analysis; Gene expression data; Regression analysis; Sign eigenanalysis; Unsupervised learning

1 Introduction

Cluster analysis, also called unsupervised learning in the study of pattern recognition, concerns the problem of the optimal partitioning of a given set of objects into a number of mutually exclusive and exhaustive clusters. It is widely applied in image processing, machine learning, taxonomy, archaeology, and the social sciences (Everitt 1993; Webb 1999). In recent years, due to the pioneering work by Eisen et al. (1998), it has become commonplace for researchers to perform cluster analysis for gene expression data to identify patterns. See Speed (2003), Satagopan and Panageas (2003) for tutorial introductions to statistical applications to gene expression data analysis.

For most applications of cluster analysis there are three important research issues to be addressed: (a) cluster discovery for the identification of clusters based on currently collected data; (b) cluster prediction by deriving a classification rule that can discriminate between the discovered clusters for new objects; and (c) variable selection for the identification of the variables which have played an important role in clustering (Satagopan and Panageas 2003; Hand, 2004).

A commonly-used class of *hierarchical* clustering methods is distance-based clustering where some heuristic definitions of the distance between objects and distance between clusters are used. The most commonly-used *non-hierarchical* clustering approach is the k-means algorithm (MacQueen, 1967). The distance-based methods and the k-means method form the backbone of cluster analysis in practice. They are widely available in software packages and easy to use (Everitt, 1993; Krzanowski and Marriot, 1995; Webb, 1999).

One of the major problems associated with these commonly-used clustering methods is the lack of a clearly defined criterion. Even if they have a clearly definited criterion, they often result in sub-optimal partitions (e.g. the k-means algorithm and Ward's algorithm, both of which aim to minimise the sum of squares). It is thus not clear how good the final partition is, and in what sense, when there is no clearly defined criterion. Further, employing different clustering methods typically yields different final partitions and one consequence is that it is not known which partition is better and thus should be adopted in practice (see, for example, the interesting discussion by Goldstein et al. (2002)). In addition, most of the commonly-used methods are purely algorithmic in the sense that they do not have any explicit model to provide a framework for further statistical analysis or to provide a link to other multivariate statistical techniques.

Besides the above difficulties in cluster discovery, these commonly-used methods do not provide a simple explicit allocation rule in terms of cluster prediction, i.e. a rule that can discriminate between clusters for new objects. Even the boundaries between clusters are given only implicitly by these methods. Finally, the problem of variable selection is very challenging for both the distance-based clustering methods (Friedman and Meulman, 2004) and for the k-means clustering algorithm (Brusco and Cradit, 2001).

In this paper, we investigate two issues of hierarchical cluster analysis, formulation and computation. First, motivated by the model of multivariate analysis of variance (MANOVA) and the method of maximum likelihood, we develop a statistical formulation of hierarchical cluster analysis based on the least squares optimization criterion, termed the clustering-function based method in this paper. Solving the formulated problem will yield a vector of group membership of objects which defines discovered clusters and a linear clustering function for classification of new objects and for variable selection.

We will also investigate the relationship between the clustering-function based method developed in this paper and model-based clustering methods. The model-based methods maximise a classification likelihood or the likelihood of a finite mixture model (Fraley and Raftery, 2002). It is shown in this paper that in theory the method of maximising classification normal likelihood and the clustering-function based method produce the same optimal partition when there are only two equal-sized groups having a common variance matrix. The actual numerical performances of the two clustering methods, however, are quite different; see Section 5 for detailed numerical comparisons.

In comparison with the k-means method and distance-based clustering methods, one major advantage of model-based clustering methods is that they have a clear criterion, maximising a likelihood. In practice, however, the model-based methods are used much less frequently than the k-means method or distance-based clustering methods. One problem is that they are likely to be trapped in a local optimum and thus numerical performance may not be as good as it appears in theory. The major reason is that cluster analysis is essentially linked to an NP problem, where the number of nontrivial partitions of N objects into g groups is $\sum_{i=1}^{g}(-1)^{g-i}\binom{g}{i}i^N/g!$ which increases exponentially with N (Everitt 1993). From a computational point of view, model-based clustering methods involve the optimisation problems which are combinatorial and/or highly nonlinear. When the number of variables is relatively large, finding a globally optimal solution for such complex problems is difficult in practice.

Because of this, it is important in cluster analysis to keep the criterion relatively simple, upon which an efficient algorithm can thus be developed. The constrained optimization problem formulated in this paper is to maximize a quadratic criterion function having both binary variables (for group membership of objects) and real-valued variables (for the coefficients of a linear clustering function). We will show analytically that a solution to this optimization problem is a sign eigenvector of a certain positive semidefinite matrix. A fast algorithm was investigated by Li (2006), exploiting the special structure of sign eigenproblems. Sign eigenanalysis provides us with a quick approach to solving the clustering problem formulated in this paper.

The paper is structured as follows. Section 2 is devoted to the motivation for the clustering-function based method. A new formulation of cluster analysis is presented in Section 3, which leads to the clustering-function based method. The issues of computation is investigated in Section 4. Simulation studies and practical examples are

examined in Sections 5 and 6. Finally, discussion and conclusions are presented in Section 7. The program written in Matlab that was used to analyse the data can be obtained from

http://www.blackwellpublishing.com/rss

2 Motivation

Hierarchical clustering proceeds by a series of successive mergers or divisions. It is argued that when the actual number of clusters is small, agglomerative methods need far more steps to reach the desired stage and may have accumulated inaccuracies by the time they reach this stage. We thus consider a divisive method in this paper.

A hierarchical divisive clustering method begins with one cluster including all objects, which is then divided into two groups such that the objects in one group are far from the objects in the other. Next, one of these groups is further divided into two dissimilar subgroups. In this paper, we use the criterion of the trace of the within-group dispersion matrix to decide which group is chosen for the next stage of division. The process continues until some stopping criterion is satisfied. Most of the stopping criteria are based on within-group dispersion and/or between-group dispersion matrices. See Everitt (1993), Krzanowski and Marriot (1995) for details.

2.1 Notation

Suppose that there are N objects measured on p variables, with observation vectors $\mathbf{x}_1, \ldots \mathbf{x}_N$, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ $(i = 1, \ldots, N)$. Assume these N objects have already been divided into several groups. The group having the largest trace of the within-group dispersion matrix is chosen to be divided further. Without loss of generality, we suppose that the chosen group consists of objects $\mathbf{x}_1, \ldots \mathbf{x}_n$ $(n < N)$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ denote the data matrix for this group. Throughout this paper we suppose that the matrix $\mathbf{X}$ is centred to zero-mean and standardised to unit variance for each of the columns. Unless stated otherwise, we also assume that matrix $\mathbf{X}$ has full rank.

For the chosen group consisting of objects $\mathbf{x}_1, \ldots \mathbf{x}_n$, we denote the vector of group membership of objects as $\mathbf{z} = [z_1, \ldots, z_n]^T$, where $\mathbf{z} \in \mathbf{Z}$, and $\mathbf{Z}$ is the space of sign vectors defined to be

$$\mathbf{Z} = \{\mathbf{z} = [z_1, \ldots, z_n]^T \mid \quad z_i = \pm 1\}.$$

All of the objects associated with an entry of $+1$ in $\mathbf{z}$ are classified into one group, whilst the others with an entry of -1 are classified into the other group.

2.2 A MANOVA model

The formulation of the clustering-function based method is motivated by the following MANOVA model with two treatments (groups) (Hoff, 2004):

$$\mathbf{x}_i = \boldsymbol{\mu} + z_i\boldsymbol{\gamma} + \boldsymbol{\epsilon}_i \quad (i = 1, ..., n), \tag{1}$$

where the error vectors $\boldsymbol{\epsilon}_i$ are assumed to be normally distributed with a zero mean and a common covariance matrix $\mathbf{V}$, i.e. $N(0, \mathbf{V})$. In addition $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_j$ $(i \neq j)$ are assumed to be independent of each other. In MANOVA, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are termed grand mean and group effect respectively. Equation (1) may be rewritten in matrix notation:

$$\mathbf{X} = \mathbf{1}\boldsymbol{\mu}^T + \mathbf{z}\boldsymbol{\gamma}^T + \boldsymbol{\epsilon},$$

where $\mathbf{1}$ is a vector of ones and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n]^T$.

2.3 Clustering problems

In this paper we treat an unsupervised learning (clustering) problem in the same way as the supervised learning (discriminant) problem except for the vector of group membership, $\mathbf{z}$, being considered as unknown. Specifically, as in Fisher linear discriminant analysis (FLDA), we consider a linear clustering function, $y = \mathbf{x}^T\boldsymbol{\beta}$, for an object $\mathbf{x}$ described by model (1), where $\boldsymbol{\beta} \neq 0$ is a vector of coefficients of the clustering function.

It is immediate from model (1) that $y_i = \mathbf{x}_i^T\boldsymbol{\beta}$ is normally distributed, $N(\boldsymbol{\mu}^T\boldsymbol{\beta} + z_i\boldsymbol{\gamma}^T\boldsymbol{\beta}, \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta})$. We reparameterise by letting $\alpha = -\boldsymbol{\mu}^T\boldsymbol{\beta}$, $\boldsymbol{\gamma}^T\boldsymbol{\beta} = 1$, and $\sigma^2 = \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}$. We thus obtain $y_i = \mathbf{x}_i^T\boldsymbol{\beta} \sim N(-\alpha + z_i, \sigma^2) \quad (i = 1, ..., n)$. Now consider a (marginal) likelihood defined by the joint distribution of y_i $(i = 1, ..., n)$:

$$L(\mathbf{z}, \alpha, \boldsymbol{\beta}, \sigma) = (2\pi)^{-n/2}\sigma^{-n}\exp\{-[\mathbf{z} - (\alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta})]^T[\mathbf{z} - (\alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta})]/(2\sigma^2)\}. \tag{2}$$

Maximising this likelihood function with respect to $\mathbf{z}, \alpha,$ and $\boldsymbol{\beta}$ is equivalent to

$$\min_{\mathbf{z}\in\mathbf{Z},\alpha,\boldsymbol{\beta}} [\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]^T[\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}],$$

which will be discussed in next section.

3 Clustering-function based method

In this section, we first develop a new clustering method, clustering-function based method, then discuss some important issues, including rank-deficiency, non-linearity, and variable selection.

For the group membership indicators z_i taking a value of either 1 or -1, consider the following linear model using a linear clustering function, $f(\mathbf{x}) = \alpha + \mathbf{x}^T\boldsymbol{\beta}$, evaluated at $\mathbf{x}_i$ to predict group membership z_i

$$z_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + e_i \quad (i = 1, ..., n), \tag{3}$$

where e_i are noise. The allocation rule is to assign $\mathbf{x}_i$ into one of the two groups according to whether $f(\mathbf{x}_i) \geq 0$ or not.

The analysis of the previous section motivates us to formulate a clustering problem as the following least squares problem:

$$\min_{\mathbf{z} \in \mathbf{Z}, \alpha, \boldsymbol{\beta}} [\mathbf{z} - \alpha \mathbf{1} - \mathbf{X}\boldsymbol{\beta}]^T [\mathbf{z} - \alpha \mathbf{1} - \mathbf{X}\boldsymbol{\beta}], \tag{4}$$

simultaneously solving for both the coefficients of the linear clustering function, α and $\boldsymbol{\beta}$, and the unknown vector of group membership, $\mathbf{z}$. According to the previous analysis, the solution to problem (4) is the estimate which maximises the likelihood (2).

For any fixed $\mathbf{z}$, it immediately follows from problem (4) that the vector of the clustering coefficients $\boldsymbol{\beta}$ is given by (since $\mathbf{X}$ is mean-centred):

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{z}. \tag{5}$$

A detailed discussion for solving problem (4) will be given in next section.

Remarks:

(i) The above formulation depends on the first two sample moments only. Hence, it is still applicable when the normal assumption in model (1) is violated, provided that the $\boldsymbol{\epsilon}_i$ have mean $\mathbf{0}$ and a common covariance matrix $\mathbf{V}$, and are mutually uncorrelated.

(ii) For one-dimensional problems, criterion (4) is equivalent to minimising the trace of within-group dispersion.

(iii) It is important to note that without constraint $\mathbf{z} \in \mathbf{Z}$, problem (4) gives a trivial solution only, i.e. $\mathbf{z} = \mathbf{0}$, $\alpha = 0$ and $\boldsymbol{\beta} = \mathbf{0}$.

3.1 Rank-deficiency or ill-conditioning design matrices

High-dimensional data analysis, for instance the analysis for gene expression data, suffers from the rank-deficiency problem, where the number of objects, n, is less than the number of variables, p. Motivated by the linear model (3), we incorporate the idea of principal-component regression to perform cluster analysis when a data matrix matrix $\mathbf{X}$ is rank-deficient or ill-conditioned. Specifically, we extract the first q $(q < min(p, n))$ principal components of matrix $\mathbf{X}$, $\mathbf{t}_1, ..., \mathbf{t}_q$, and rewrite the linear model (3) as:

$$\mathbf{z} = \alpha \mathbf{1} + \sum_{j=1}^{q} \theta_j \mathbf{t}_j + \tilde{\mathbf{e}}, \tag{6}$$

where $\tilde{\mathbf{e}}$ is a noise vector, including both the noise $\mathbf{e}$ in equation (3) and the residual of $\mathbf{X}$ after extracting the first q principal components. The θ_j are scalar coefficients. Cluster analysis can thus be carried out based on equation (6).

The remaining issue is how to determine the number of principal components, q. Let $\mathbf{W}$ and $\mathbf{B}$ represent the within-group and between-group dispersion matrices respectively. In this paper we suggest a simple criterion: select q such that $trace(\mathbf{B})/trace(\mathbf{W})$ is maximised.

3.2 Non-linear clustering

Many existing clustering methods are likely to be successful when clusters are elliptical (Everitt, 1993, chapter 5). This is also true for the linear clustering function, $f(\mathbf{x}) = \alpha + \mathbf{x}^T\boldsymbol{\beta}$, developed in this paper. The linear clustering function results in a straight-line or hyperplane boundary between two groups.

In practice, however, clusters can have much more complicated shapes, and boundaries between clusters can be very complex curves or surfaces. To deal with this problem, we incorporate the idea of polynomial regression or extended linear models via B-splines. For instance, the linear model (3) may be extended to the following quadratic model:

$$z_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j,k=1}^{p} x_{ij}x_{ik}\beta_{jk} + e_i \quad (i = 1, ..., n). \tag{7}$$

Such extensions provide a considerable flexibility when dealing with non-linear boundaries bewteen groups.

3.3 Variable selection

Without an appropriate model, variable selection would be a very difficult issue in cluster analysis. The recent research on variable selection in cluster analysis by Friedman and Meulman (2004) shows how difficult it can be.

Under the formulation developed in this paper, variable selection can be carried out informally by simply inspecting the relative magnitudes of the standardised coefficients of a clustering function. Those variables having relatively small magnitudes of coefficients are regarded as less important for clustering and may thus be removed from the clustering function if the reduced clustering function results in the same partition.

Alternatively, since the formulation developed in this paper is linked to discriminant analysis (see next subsection), variable selection can be carried out using various variable selection methods in discriminant analysis, conditional on the group membership vector $\mathbf{z}$.

3.4 Relationship with FLDA

The major difference between the above formulation of cluster analysis and that of FLDA is that the group membership vector $\mathbf{z}$ is known *a priori* in FLDA. Denote the population means of two groups in FLDA as $E(\mathbf{x}|z=1)$ and $E(\mathbf{x}|z=-1)$ respectively. Let $\boldsymbol{\Delta} = [E(\mathbf{x}|z=1) - E(\mathbf{x}|z=-1)][E(\mathbf{x}|z=1) - E(\mathbf{x}|z=-1)]^T$ and denote the common covariance matrix as $\mathbf{V}$. In FLDA, a linear discriminant function $\mathbf{x}^T\boldsymbol{\beta}$ is determined by (Krzanowski and Marriot, 1995):

$$\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta}}{\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}}.$$

In cluster analysis we may use the same criterion as above and seek a linear clustering function $\mathbf{x}^T\boldsymbol{\beta}$ *and* an *unknown* membership vector $\mathbf{z}$ simultaneously:

$$\max_{\mathbf{z}\in\mathbf{Z},\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^T\boldsymbol{\Delta}\boldsymbol{\beta}}{\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}}. \tag{8}$$

Note that from equation (1) we have $E(\mathbf{x}|z=1) - E(\mathbf{x}|z=-1) = 2\boldsymbol{\gamma}$, and thus $\boldsymbol{\Delta} = 4\boldsymbol{\gamma}\boldsymbol{\gamma}^T$. In practice, when population parameters are not available they are often replaced by their sample counterparts. For this particular problem, the maximum likelihood estimate of the common covariance matrix $\mathbf{V}$ is used for fixed $\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, and $\mathbf{z}$:

$$\hat{\mathbf{V}} = [\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{z}\boldsymbol{\gamma}^T]^T[\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{z}\boldsymbol{\gamma}^T]/n.$$

Hence, following the same reparameterisation, $\alpha = -\boldsymbol{\mu}^T\boldsymbol{\beta}$ and $\boldsymbol{\gamma}^T\boldsymbol{\beta} = 1$, the optimization problem (8) reduces to problem (4) by replacing $\mathbf{V}$ by $\hat{\mathbf{V}}$.

3.5 Link to model-based clustering methods

Consider a population consisting of g different subpopulations, each having a density function $f_k(\mathbf{x};\boldsymbol{\phi})$ with parameter vector $\boldsymbol{\phi}$ $(k = 1, ..., g)$. Given objects $\mathbf{x}_1$,...,$\mathbf{x}_N$, let $\mathbf{z} = [z_1, ..., z_N]^T$ denote the group membership indicators, where $z_i = k$ if $\mathbf{x}_i$ comes from the kth subpopulation. There are two types of model-based clustering methods. The first one is to choose parameters $\boldsymbol{\phi}$ and partition $\mathbf{z}$ so as to maximise the classification likelihood $L_{CL}(\boldsymbol{\phi}, \mathbf{z}) = \prod_{i=1}^{N} f_{z_i}(\mathbf{x}_i;\boldsymbol{\phi})$. The second approach is to maximise the likelihood of a mixture of densities $f_k(\mathbf{x};\boldsymbol{\phi})$,

$$L_{MIX}(\boldsymbol{\phi}, \pi_1, ..., \pi_g) = \prod_{i=1}^{N}\sum_{k=1}^{g} \pi_k f_k(\mathbf{x}_i;\boldsymbol{\phi}),$$

where π_k are unknown mixing proportions (Fraley and Raftery, 2002). It is shown by Celeux and Govaert (1993) that, numerically, neither of the two model-based clustering approaches is uniformly superior to the other.

When $f_k(\mathbf{x};\boldsymbol{\phi})$ is multivariate normal with mean vector $\boldsymbol{\xi}_k$ and covariance matrix $\mathbf{V}_k$, Banfield and Raftery (1993) demonstrate that the criterion of maximising the

classification likelihood reduces to: (a) the Ward's criterion (1963), i.e. $\min \text{trace}(\mathbf{W})$, if $\mathbf{V}_k = \sigma^2\mathbf{I}$ for all i, where $\mathbf{I}$ is an identity matrix; (b) the Friedman-Rubin criterion (1967), i.e. $\min \det(\mathbf{W})$, if $\mathbf{V}_k = \mathbf{V}$ for all i, where $\mathbf{V}$ is a common covariance matrix.

The lemma below shows that the clustering-function based method is closely related to the method of maximising the classification likelihood based on normal component densities.

Lemma 1 *For two-group clustering problems, if two groups have equal size and a common covariance matrix, then maximising the classification likelihood $L_{CL}(\boldsymbol{\phi}, \mathbf{z})$ based on normal component densities and minimising the least squares criterion (4) produce the same optimal partition.*

See the Appendix for proof. Lemma 1 provides another motivation for the clustering-function based method. It is clear from Lemma 1 that if a linear clustering function $f(\mathbf{x}) = \alpha + \mathbf{x}^T\boldsymbol{\beta}$ is imposed, the method of maximising the classification likelihood based on normal component densities implicitly defines the coefficients of the linear clustering function via the least square solution (5).

4 Solutions and computation

In this section, we solve problem (4) analytically and investigate the issues of computation. First we consider a simpler case where a set of objects in cluster analysis are to be divided into two groups with approximately equal sizes. The general situation will be investigated later.

4.1 Partition into two approximately equal-sized groups

When two groups to be divided have approximately equal sizes, the mean of partition $\mathbf{z}$, $\bar{\mathbf{z}} = \sum_{i=1}^{n} z_i/n$, is approximately zero. There is thus no need for the intercept in equation (3). The problem (4) reduces to

$$\min_{\mathbf{z}\in\mathbf{Z},\boldsymbol{\beta}} [\mathbf{z} - \mathbf{X}\boldsymbol{\beta}]^T[\mathbf{z} - \mathbf{X}\boldsymbol{\beta}]. \tag{9}$$

As mentioned earlier, the computation of the clustering-function based method will be converted to that of sign eigenanalysis. For this end, we first define a sign function, $S(x)$, to be

$$S(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ -1 & \text{if} \quad x < 0. \end{cases}$$

When $\mathbf{x}$ is a vector, $S(\mathbf{x})$ is a vector of the same dimension as $\mathbf{x}$ containing the signs of the elements of $\mathbf{x}$. A sign eigenvector of an $n \times n$ positive semidefinite matrix $\mathbf{A} \geq 0$ is defined to be a sign vector $\mathbf{z} \in \mathbf{Z}$ satisfying (Li, 2006)

$$\mathbf{z} = S(\mathbf{A}\mathbf{z}),$$

where the associated sign eigenvalue is defined to be $\mathbf{z}^T\mathbf{A}\mathbf{z}/n$. According to this definition, a sign eigenvector of a matrix $\mathbf{A} \geq 0$ is a sign vector for which all of its entries retain the same signs after the linear transformation $\mathbf{A}$.

Define $\mathbf{H}_c = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ for linear model (3). The main result for solving problem (9) is summarised in the following theorem.

Theorem 1 *For a full rank and mean-centered data matrix* $\mathbf{X}$*, a solution to problem (9) is given by* $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$*, where* $\mathbf{z}$ *is a sign eigenvector of* $\mathbf{H}_c$ *associated with the largest sign eigenvalue, satisfying* $\mathbf{z} = S(\mathbf{H}_c\mathbf{z})$*.*

The proof is given in the Appendix. For a positive semidefinite matrix, it is shown by Li (2006) that there always exists at least one sign eigenvector, which thus guarantees the existence of a solution to problem (9). Like ordinary eigenvectors, sign eigenvectors associated with the largest sign eigenvalue may not be unique. When this occurs, although it is rare for real-world data, we have to accept all of the resultant partitions or select one particular partition using some other criterion.

To illustrate the above theorem, we consider a simple example below.

Example 4.1. Consider a clustering problem for the data set $\{1, 2, 5, 7, 9, 10\}$. The aim is to divide this data set into two approximately equal-sized groups. For this one-dimensional problem we may work out the optimal solution straightforwardly. The standardised data matrix is given by $\mathbf{X} = [-1.272, -0.999, -0.182, 0.363, 0.908, 1.181]^T$. The only sign eigenvectors of $\mathbf{H}_c$ are $\mathbf{z} = \pm[-1, -1, -1, 1, 1, 1]^T$. The two resultant clusters are thus given by $\{1, 2, 5\}$ and $\{7, 9, 10\}$ with a sum of squares of 13.33. Note that this is the minimum of the sum of squares. We may draw a comparison with Ward's algorithm (1963). Ward's algorithm is an agglomerative hierarchical clustering algorithm where the criterion is to minimise the sum of squares. Ward (1963) developed a fast algorithm to solve clustering problems although it cannot guarantee to achieve a global optimum. For the data set $\{1, 2, 5, 7, 9, 10\}$, Ward's algorithm results in two groups, $\{1, 2\}$ and $\{5, 7, 9, 10\}$, with a sum of squares of 15.25. Hence this is a sub-optimal partition.

Next, we discuss computational issues. Let $\mathbf{A} = \mathbf{H}_c$. According to Li (2006), a sign eigenvector of $\mathbf{A} \geq 0$ associated with the largest sign eigenvalue is a solution to $\max_{\mathbf{z}\in\mathbf{Z}} \mathbf{z}^T\mathbf{A}\mathbf{z}$. An alternating algorithm for solving the sign eigenproblem was investigated by Li (2006). It incurs very low computational costs and can be easily implemented:

ALGORITHM 1.

Step 1. INITIALISATION. Set an initial sign vector $\mathbf{z}_0$.

Step 2. REPEAT

LET $\mathbf{z}_k = S(\mathbf{A}\mathbf{z}_{k-1})$ for $k = 1, 2, 3, ...,$

UNTIL $\mathbf{z}_k = \mathbf{z}_{k-1}$.

END.

Note that like most other nonlinear programming techniques, this algorithm may be trapped in a local optimum. Multiple initial guesses $\mathbf{z}_0$ have to be tried. In addition, this strategy can be combined with a k-depth perturbation scheme for a pre-selected k (say $k = 5$), where whenever a (local) maximum is attained at, say $\mathbf{z}$, we change every collection of j ($j = 1, ..., k$) entries of $\mathbf{z}$ to their opposite signs. If the objective function, $\mathbf{z}^T\mathbf{A}\mathbf{z}$, is improved at any of the modified $\mathbf{z}$, jump to this new solution and call Algorithm 1 once again using the new solution as initial guess; otherwise, stop.

4.2 Partition in the general situation

Next, we consider solving problem (4) under the general situation where objects are to be divided into two groups, not necessarily equal-sized.

Define $\mathbf{H}(\tau) = \mathbf{H}_c + (\tau/n)\mathbf{1}\mathbf{1}^T$. From Theorem 1, we may obtain the following solution: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$ and $\alpha = \bar{\mathbf{z}}$, where $\mathbf{z}$ is a sign eigenvector of $\mathbf{H}(1)$ associated with the largest sign eigenvalue, satisfying $\mathbf{z} = S(\mathbf{H}(1)\mathbf{z})$.

Lemma 2 *The n-vectors $\mathbf{1}$ and $-\mathbf{1}$ are sign eigenvectors of matrix $\mathbf{H}(1)$ associated with the largest sign eigenvalue, 1.*

The proof of Lemma 2 is given in the Appendix. According to Lemma 2, unless there exists some other sign eigenvector of $\mathbf{H}(1)$ associated with the largest sign eigenvalue, problem (4) is not a well-defined problem since it produces a trivial partition where all objects are classified into a single group (since the entries of the resultant partition $\mathbf{z} = \pm\mathbf{1}$ are all ones or all minus ones), $\boldsymbol{\beta} = \mathbf{0}$ and α is either 1 or -1.

To obtain a non-trivial solution, we have to exclude the single-group partition, i.e. $\mathbf{z} = \mathbf{1}$ and $\mathbf{z} = -\mathbf{1}$. This is equivalent to enforcing a constraint of non-zero coefficients on the linear clustering function, $||\boldsymbol{\beta}||_{\mathbf{M}} \neq 0$, where $\mathbf{M}$ is a metric matrix. In this paper we choose a metric matrix $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ which is a commonly-used choice in regression analysis (see e.g., Weisberg, 1985). Mathematically, we consider the following problem:

$$\min_{\mathbf{z}\in\mathbf{Z},\alpha,\boldsymbol{\beta}} \frac{[\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]^T[\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]}{\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}}. \tag{10}$$

The solution to this problem is summarised in the following theorem. The proof is given in the Appendix.

Theorem 2 *For a full rank and mean-centered data matrix $\mathbf{X}$, an optimal solution to problem (10) is given by $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}/(1-\lambda)$, $\alpha = \bar{\mathbf{z}}$, and $\mathbf{z}$ is a sign eigenvector*

of $\mathbf{H}(1-\lambda)$ *associated with the largest sign eigenvalue, satisfying* $\mathbf{z} = S(\mathbf{H}(1-\lambda)\mathbf{z})$, *where* $\mathbf{H}(\tau) = \mathbf{H}_c + (\tau/n)\mathbf{1}\mathbf{1}^T$ *and* $\lambda = [\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]^T[\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]/(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})$.

Note that from Theorem 2, an established linear clustering function may be equivalently written as

$$f(\mathbf{x}) = \hat{\alpha} + \mathbf{x}^T\hat{\boldsymbol{\beta}}, \tag{11}$$

with $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$ and $\hat{\alpha} = (1-\lambda)\overline{\mathbf{z}}$.

We may interpret the above result as follows. Matrix $\mathbf{H}(1)$ is perturbed as little as possible such that the perturbed matrix $\mathbf{H}(1-\lambda)$ has a sign eigenvector associated with the largest sign eigenvalue which differs from both $\mathbf{1}$ and $-\mathbf{1}$. Theorem 2 guarantees such perturbation is minimum in terms of λ .

From Theorem 2, it is immediate to have the following result.

Corollary 1 *If an optimal partition* $\mathbf{z}^*$ *to problem (10) divides objects into two equal-sized clusters, i.e.* $\overline{\mathbf{z}}^* = 0$, *then it is also an optimal partition to problem (9).*

This corollary indicates that Theorem 1 for equal-sized problems in Section 4.1 is a special case of Theorem 2.

Now we consider computational issues. Since for any given λ, we may calculate a sign eigenvector $\mathbf{z}$ of $\mathbf{H}(1-\lambda)$ associated with the largest sign eigenvalue, the remaining issue is how to find a suitable value of λ. For this end, we first state a lemma.

Lemma 3 *The optimal objective value of (10),* λ^*, *associated with the optimal partition* $\mathbf{z}^*$ *is equal to* $\lambda^* = 1 - (\mathbf{z}_c^{*T}\mathbf{H}_c\mathbf{z}_c^*)/(\mathbf{z}_c^{*T}\mathbf{z}_c^*)$, *where* $\mathbf{z}_c^* = \mathbf{z}^* - \overline{\mathbf{z}}^*\mathbf{1}$.

The proof of the lemma is immediate from Theorem 2. Since λ^* lies between 0 and 1, from Lemma 3 we may use a linear search on the interval $[0, 1]$ for the optimal value of λ, starting from 0.

ALGORITHM 2.

Step 1. INITIALISATION. Set $\tau = 1$ and step $r \in (0, 1)$.

Step 2. REPEAT

LET $\tau = \tau - r$ and $\mathbf{H}(\tau) = \mathbf{H}_c + (\tau/n)\mathbf{1}\mathbf{1}^T$

CALL Algorithm 1 for a sign eigenvector $\mathbf{z}$ of $\mathbf{H}(\tau)$ associated with the largest sign eigenvalue

UNTIL $\mathbf{z} \neq \pm\mathbf{1}$ or $\tau \leq 0$.

END.

Note that if Algorithm 2 terminates with $\tau \leq 0$, the step r has to be replaced by a smaller one. The following example illustrates this algorithm.

Example 4.2. Consider a clustering problem for data set $\{1, 2, 5, 7, 9, 16\}$. Intuitively one object, $\{16\}$, lies far away from the remaining objects. We wish to divide this data set into two groups, hopefully one group having a single object $\{16\}$ and the other group consisting of $\{1, 2, 5, 7, 9\}$. When solving problem (10), we find that for $\lambda \leq 0.30$, the perturbed matrix $\mathbf{H}(1-\lambda)$ has sign eigenvectors $\mathbf{1}$ and $-\mathbf{1}$ only, both being trivial solutions. As the iteration process continuous such that λ becomes slightly greater than 0.30, for instance, lying between 0.30 and 0.34, matrix $\mathbf{H}(1-\lambda)$ has sign eigenvectors $\pm[1, 1, 1, 1, 1, -1]^T$ associated with the largest sign eigenvalue. This results in the optimal partition with two clusters, $\{1, 2, 5, 7, 9\}$ and $\{16\}$.

In the above example we have seen that any value of λ between 0.30 and 0.34 will give the desired partition. In general, since the entries of partition $\mathbf{z}$ take values ± 1 only, an optimal partition is typically not very sensitive to λ. Hence, the step r in Algorithm 2 does not have to be set too small in practice. Once a desired partition is obtained, a suitable optimal objective value λ^* may be calculated from Lemma 3 to form a clustering function (11).

5 Simulation studies

In this section, the general approach of the clustering-function based method developed by Theorem 2 is illustrated using simulated data. For comparison, some commonly-used clustering methods are also applied using SPSS12.0. We also consider model-based clustering methods. Since neither the method of finite mixture models nor the method of classification likelihood has uniformly superior performance over the other (Celeux and Govaert, 1993), we consider the finite mixture model only. The finite mixture model considered here is assumed to have normal component densities with unequal covariance matrices. The EM algorithm is used for maximising the mixture likelihood. Following Celeux and Govaert's (1993) strategy for addressing the problem of finding a suitable starting value, we carry out 100 runs for each of the examples with randomly selected initial guess and the best likelihood-based solution is taken as the resultant partition of the finite mixture model.

5.1 Variable selection

Example 5.1. Three groups of data measured on $p = 20$ variables were simulated using normal distributions, $N(\boldsymbol{\xi}_i, \mathbf{V}_i)$ $(i = 1, 2, 3)$, where $\boldsymbol{\xi}_1 = [1.5, -1.5, \mathbf{0}_{1\times(p-2)}]^T$, $\boldsymbol{\xi}_2 = \mathbf{0}_{p\times 1}$ and $\boldsymbol{\xi}_3 = [3, 1.5, \mathbf{0}_{1\times(p-2)}]^T$. The covariance matrices are

$$\mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{0i} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{18\times 18} \end{pmatrix} \quad \text{where} \quad \mathbf{V}_{0i} = \begin{pmatrix} 0.25 & 0.1875 r_i \\ 0.1875 r_i & 0.25 \end{pmatrix},$$

where $r_1 = r_2 = 1$ and $r_3 = -1$. $\mathbf{I}$ is an identity matrix. The population distributions generating the three groups differ only in the first two variables. Each group consists of

100 objects. We first used the k-means algorithm and some distance-based algorithms to perform cluster analysis. The results are displayed in Table 1. For example, when using the k-means algorithm with the number of groups specified as 3, out of the 100 objects in group 1 there were 35 misclassified into group 2, and 12 misclassified into group 3.

Table 1. Cluster analysis using existing clustering methods

From groups	Number of objects classified into groups			
	I	II	III	Total
I	$53^{\#}$ $(70)^{+}$ [62]* $\{40\}^{\star}$	35 (24) [34] {39}	12 (6) [4] {21}	100
II	33 (59) [10] {38}	55 (38) [81] {47}	12 (3) [9] {15}	100
III	8 (24) [59] {46}	4 (8) [35] {22}	88 (68) [6] {32}	100
Total	94 (153) [131] {124}	94 (70) [150] {108}	112 (77) [19] {68}	

$^{\#}$Figures are classification by the k-means algorithm

$^{+}$Figures in parentheses are classification by the Ward's algorithm

*Figures in brackets are classification by the complete linkage algorithm

$\star$ Figures in braces are classification by the finite mixture model

The overall misclassification numbers are 104 for the k-means algorithm, 124 for the Ward's algorithm, 151 for the complete linkage algorithm, and 181 for the finite mixture model. Note that the data were generated under the assumption of a finite normal mixture model but the partition obtained via this model is far from the true groupings.

Cluster analysis was also performed using the clustering-function based method. During the first stage, the 300 objects were classified into two groups. In the second stage, the group having a larger trace of within-group dispersion was selected and the clustering-function based method was applied to this chosen group. All of the 300 objects were correctly classified. The resultant vector $\boldsymbol{\beta}$ of the standardised coefficients of the linear clustering function for the first stage is:

$$[0.527, 0.512, -0.001, -0.004, 0.014, -0.011, -0.031, 0.011, 0.018, 0.009,$$
$$0.009, 0.003, 0.014, 0.005, 0.019, -0.013, 0.005, -0.005, 0.009, -0.005]^T$$

with an estimate of intercept α of -0.298, and the coefficient vector for the second stage clustering fucntion is

$$[-0.532, 0.592, 0.018, 0.013, 0.033, 0.001, -0.017, -.021, -0.008, -0.016,$$
$$0.016, -0.014, -0.017, 0.033, 0.008, -0.014, 0.017, -0.021, -0.004, -0.009]^T$$

with an estimate of intercept α of 0.

Clearly except for the first two entries, all the remaining entries of the above two vectors are very small, indicating that the corresponding variables are less important for both stages. This can be verified via a formal variable selection procedure in discriminant analysis, conditional on the resultant vectors of group membership.

Based on the above analysis, we kept the first two variables only for both stages of clustering. Re-performing cluster analysis yielded clustering functions, $z = -0.297 +$

$0.530x_1 + 0.510x_2$ for the first stage and $z = -0.537x_1 + 0.586x_2$ for the second stage. Once again, this produced a perfect partition with no misclassification.

This example indicates that commonly-used clustering algorithms are likely to be influenced by the variables which are irrelevant to clustering. Using the clustering-function based method, however, it is easy to identify irrelevant variables and thus remove them from cluster analysis.

5.2 Non-linear boundaries between groups

Example 5.2. Two groups of data measured on $p = 3$ variables, x_1, x_2, and x_3, were simulated. The first group consists of 80 objects, as displayed in Fig. 1 with marker '*', and the second group consists of 120 objects with marker '+'. For group 1, $x_1 \sim U(-3\pi/2, \pi)$ and $x_2 = 2\sin x_1^2 + e_1$, whilst for group 2, $x_1 \sim U(-\pi, 3\pi/2)$ and $x_2 = 2\sin x_1^2 + e_2$, where $e_1 \sim U(0, 2)$ and $e_2 \sim U(0, 2)$. All of these uniform variates are independent of each other. Variate x_3 is independent of x_1 and x_2 and has $N(0, 1)$ in both groups.

Cluster analysis was performed using the clustering-function based method with a cubic clustering function but the interaction terms were included only up to the second order. Using this method, all of the 200 objects were correctly classified. The calculated boundary (dashed line) in the x_1x_2-plane is displayed in Fig. 1.

For comparison, the k-means method with the number of groups specified as 2 was also applied. The misclassified objects by the k-means algorithm are circled in Fig. 1. In addition, we also performed cluster analysis using other existing methods. The results are displayed in Table 2. The overall misclassification numbers are 68 for the k-means algorithm, 49 for the Ward's algorithm, 90 for the complete linkage algorithm, and 36 for the finite mixture model.

Table 2. Cluster analysis using existing clustering methods

From groups	Number of objects classified into groups		
	I	II	Total
I	$56^{\#}$ (66)$^{+}$ [80]* {67}*	24 (14) [0] {13}	80
II	44 (35) [90] {23}	76 (85) [30] {97}	120
Total	100 (101) [170] {90}	100 (99) [30] {110}	

$^{\#}$Figures are classification by the k-means algorithm

$^{+}$Figures in parentheses are classification by the Ward's algorithm

*Figures in brackets are classification by the complete linkage algorithm

$\star$ Figures in braces are classification by the finite mixture model

Before concluding this section, it should be mentioned that the classification results of the clustering-function based method displayed in this section represent best case performance. For other simulated data sets, the performance of the clustering-function based method may differ.

6 Practical examples

In this section we examine some practical examples. For ease of interpretation and comparison, we choose data sets for which the group membership of objects is already known. The information about group membership is not used during clustering; it is used solely for the evaluation of the performances of the clustering methods.

6.1 The Fisher iris data

The Fisher iris data were measured on four variables, sepal length, sepal width, petal length, and petal width, in millimeters. There were 50 iris specimens from each of three species, setosa, versicolor, and virginica. The purpose of the analysis here is to discover three clusters associated with the three species based solely on the measurements.

6.1.1 Cluster analysis using existing methods

The analysis of Fisher iris data using the Ward's algorithm and the k-means clustering algorithm are provided in SAS-STAT user's guide (1999) and the results are displayed below. It can be seen that they have almost the same performances for this data set.

Table 3. Cluster analysis using existing methods for Iris data

From species	Number of objects classified into species			
	setosa	versicolor	virginica	Total
setosa	$50^{\#}$ $(50)^{+}$	0 (0)	0 (0)	50
versicolor	0 (0)	48 (49)	2 (1)	50
virginica	0 (0)	14 (15)	36 (35)	50
Total	50 (50)	62 (64)	38 (36)	

$^{\#}$Figures are classification by the k-means algorithm

$^{+}$Figures in parentheses are classification by the Ward's algorithm

The overall misclassification numbers are 16 for the k-means algorithm and 16 for the Ward's algorithm.

6.1.2 Cluster analysis using the clustering-function based method

Next, cluster analysis was performed using the clustering-function based method developed by Theorem 2. The first stage of cluster analysis yielded two groups, the first one consisting of all 50 objects of setosa, and the second including the remaining 100 objects belonging to versicolor and virginica groups. Carrying out the second-stage analysis for the second group which had a larger trace of within-group dispersion, the second group of 100 objects was divided into two sub-groups, where 2 out of the 50 objects in versicolor group were misclassified into virginica group, and 1 out of the 50 objects

in virginica group was misclassified into versicolor group. The overall misclassification number is 3.

To see what this optimal partition means to us, we used the information of the known membership (species of the 150 observations) to perform FLDA. This resulted in the same classification as the clustering-function based method. This indicates that for this data set, the classification cannot be improved even if we have extra information about the group membership. In other words, supervised learning and unsurprised learning produce the same classification results for this data set. This can occur in cluster analysis if objects are suitably separated. To put it another way, supervison in learning process is not necessary if objects in different groups are characterised by some certain features.

6.2 Prostate cancer gene expression data

LaTulippe et al. (2002) carried out a gene expression data analysis of prostate cancer. The data set consisted of gene expressions of 32 cancer patients: tissues from the primary disease site of 23 patients with no known metastasis at the time of tissue extraction, and tissues from the metastatic disease site of 9 patients. Expression data on approximately 63000 genomic regions were obtained using five oligonucleotide chips for every patient. Satagopan and Panageas (2003) analysed the data from only one chip, which included 12626 genes. The aims of their analysis were to identify diseases based on the patient's gene expression without utilising any prior knowledge of the groups, and to identify genes that can discriminate between primary and metastatic groups. Ideally, the group of 23 primary samples can be separated from the group of 9 metastatic samples.

6.2.1 Cluster analysis using existing methods

Satagopan and Panageas (2003) first carried out cluster analysis using the distance-based method with average linkage and Pearson correlation coefficients as the distance between objects. As Satagopan and Panageas (2003) noted, it appeared that two general clusters could be identified, one consisting of 8 primary samples, and the other consisting of the remaining 24 samples, as shown in Table 4.

Table 4. Cluster analysis of prostate cancer gene expression data by the average linkage algorithm and the k-means algorithm

From diseases	Number of samples classified into diseases		
	cluster I	cluster II	Total
primary	$8^{\#}$ (15)*	15 (8)	23
metastasis	0 (0)	9 (9)	9
Total	8 (15)	24 (17)	

#Figures are classification by the average linkage algorithm
*Figures in parentheses are classification by the k-means algorithm

Satagopan and Panageas (2003) also considered the k-means algorithm. For the number of clusters specified as 2, the k-means algorithm yielded a better partition as shown in Table 4. Furthermore, Satagopan and Panageas (2003) investigated the issue of choosing the number of clusters by using F-values. It was shown that for the k-means algorithm, the appropriate number of clusters was 2 for this data set.

6.2.2 Cluster analysis using the clustering-function based method

Next, we use the clustering-function based method developed by Theorem 2 to analyse this data set. The following three issues will be addressed: (a) cluster discovery; (b) variable identification/selection; and (c) cluster prediction.

Discovery of clusters

The principal-component based model (6) was used to establish a clustering function since the number of the variables (genes), $p = 12626$, was larger than the number of the samples, $n = 32$. The criterion discussed in Section 3.1 suggested that the number of components, q, should be 2 to 4 (see Fig. 2). When $q = 4$ was taken, the derived clustering function yielded a perfect partition with no misclassification: all primary samples were classified into one group, whilst all the metastatic samples were classified into the other. The resulting clustering function is

$$z = -0.380 + 0.205t_1 + 0.822t_2 - 0.052t_3 - 0.070t_4, \tag{12}$$

where t_i is the ith principal component.

At this stage it is interesting to raise a question: can the classifications be improved if we use the k-means method and the average linkage algorithm again to analyse the 'cleaned' data, i.e. the first four principal components? When the number of clusters was specified as 2, applying the k-means method yielded a partition where one group consisted of only two samples, number 28 and number 31, whilst the other group consisted of all the remaining samples. Applying the average linkage algorithm yielded a partition where one group consisted of only one sample, number 31, whilst the other group consisted of all the remaining samples. Hence, for this example, classifications cannot be improved by these methods using the 'cleaned' data.

Returning to the obtained clustering function (12), we noted that the standardised coefficients of t_3 and t_4 were relatively small, suggesting that t_3 and t_4 should be relatively less important. We thus removed these two components from the clustering function (12) and carried out cluster analysis once again using the first two principal components only. This again yielded a perfect partition with no misclassification.

Variable identification

To identify genes that can discriminate between primary and metastatic groups without utilising any prior knowledge about the groups, we considered the loadings of the first two components. For each of these two components, we picked those genes where the absolute values of the loadings were greater than 0.05. In total 18 genes were selected. We then performed cluster analysis using these 18 genes only. This again yielded a perfect partition with no misclassification. However, 13 of the 18 genes had relatively small standardised clustering coefficients and thus were removed from the clustering function. This left only five genes, i.e. genes $3236, 5145, 6033, 7167$, and 11962. Cluster analysis was performed again using these five genes, and again yielded a perfect partition with no misclassification. The resulting clustering function is

$$z = -0.437 + 0.109x_{3236} - 0.138x_{5145} - 0.175x_{6033} - 0.836x_{7167} + 0.139x_{11962}.$$

It is somewhat amazing that of the original 12626 genes, only 5 genes are sufficient to discriminate between primary and metastatic groups without utilising any prior knowledge about the groups.

Cluster prediction

To address the issue of cluster prediction, the leave-one-out method was used for the evaluation of the performance of the clustering-function based method. This is a commonly-used method in discriminant analysis (Krzanowski and Marriot, 1995). Specifically, each of the 32 samples was omitted in turn from the data, and the remaining samples were used to build a clustering function. The obtained clustering function was then employed to predict the group membership of the omitted sample. When the above five genes were used for cluster analysis, the leave-one-out error rate was zero, i.e. all of the group membership of the omitted samples were correctly predicted.

6.3 Melanoma cancer gene expression data

Next, we briefly discuss the gene expression data of cutaneous malignant melanoma cancer analysed by Bittner et al. (2000). By using the average linkage algorithm, Bittner et al. (2000) suggested a partition of two clusters, where cluster I consisted of samples 13 to 31. The remaining samples, i.e. samples 1-12, although scattered into several separate groups in the dendrogram (not shown here, see Fig. 1 in Goldstein et al., 2002), were considered as cluster II by Bittner et al. (2000). Goldstein et al. (2002) analysed the same gene expression data using several other distance-based clustering methods. They questioned the reliability of the discovery by Bittner et al. (2000), arguing that their clustering result was not re-producible when the other commonly-used linkages were considered.

To analyse the data of Bittner et al. (2000), we used model (6) to build a clustering function. The criterion in Section 3.1 suggested $q = 4$. The derived clustering function yielded a two-cluster partition where the first cluster consisted of sample 1 and samples

13-31, and the second cluster consisted of samples 2-12. Interestingly, although the discovered pattern by Bittner et al. (2000) was not recognisable by most of the other distance-based methods (Goldstein et al., 2002), our derived clustering function yielded a partition with only one sample (i.e. sample 1) differing from the discovered pattern by Bittner et al. (2000), thus it seems to support their discovery. The obtained clustering function is

$$z = -0.257 + 0.211t_1 + 0.414t_2 + 0.292t_3 + 0.733t_4,$$

where t_i is the ith principal component.

7 Discussion and conclusions

In this paper, a clustering-function based method is developed for cluster analysis. It provides a nice link to discriminant analysis, regression analysis, and some existing clustering methods. Consequently, it is relatively easy and straightforward to handle some difficult issues such as high-dimensional data, non-linearity and variable selection which it is hard to deal with using the existing clustering methods. Numerical examples show that the proposed clustering method outperforms existing clustering methods.

The clustering-function based method developed in this paper has some limitations. First, as a hierarchical method, it has all problems associated with hierarchical clustering. Secondly, its close link to FLDA seems to imply that it would probably fail in any situation where FLDA fails. Knoke (1982) constructs examples where FLDA has very poor performance when continuous variables and binary variables interact in certain ways.

In general, however, various comparative studies have shown that FLDA is quite robust (see, e.g. Knoke, 1982; Krzanowski and Marriot, 1995). The performance of FLDA can be further improved when it is suitably extended, such as by including higher-order terms or building an appropriate location model (Krzanowski, 1975). Such extensions also make sense for the clustering method proposed in this paper.

The motivation of the clustering-function based method is based on equation (1), assuming normal densities with a common covariance matrix. It is obvious that such assumptions are quite likely to be violated in practice. However, as with FLDA, the proposed method essentially depends on the first two sample moments only via equation (4). The numerical analysis discussed in this paper shows that the proposed method is not sensitive to the assumption of either normal densities or common covariance matrix. Nevertheless, further study to investigate how robust this method is would be desirable.

Acknowledgements

The author is very grateful to the Associate Editor and two referees for their valuable comments which have improved this paper. He also wishes to thank Dr. J. M. Satagopan for providing prostate cancer gene expression data, Prof. D. Ghosh for providing melanoma cancer gene expression data, and his colleague, Dr. D. Coates, for his helpful comments.

Appendix: proofs of theorems

A.1 Proof of Theorem 1

Since $\mathbf{z}^T\mathbf{z} = n$, the criterion function in equation (9) may be rewritten as

$$\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} - 2\mathbf{z}^T\mathbf{X}\boldsymbol{\beta} + n.$$

Differentiating the above criterion function with respect to $\boldsymbol{\beta}$ and equating it to zero yield the solution $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}$. In addition, by applying Theorem 2 in Li (2006), we establish that $\mathbf{z}$ is a sign eigenvector of $\mathbf{H}_c$ associated with the largest sign eigenvalue. This completes the proof.

A.2 Proof of Lemma 1

For the case of common covariance matrix, maximising a classification likelihood with normal component densities is equivalent to solving $\min \det(\mathbf{W})$ (Banfield and Raftery, 1993). For two-group problems, the group labels z_i may be recoded as 1 and -1. Therefore, when the two groups have equal size and a common covariance matrix, the within-group dispersion matrix is $\mathbf{W} = [\mathbf{X} - \mathbf{z}\mathbf{z}^T\mathbf{X}/n]^T[\mathbf{X} - \mathbf{z}\mathbf{z}^T\mathbf{X}/n]$ and

$$\det(\mathbf{W}) = \det(\mathbf{X}^T\mathbf{X})[\mathbf{I} - \mathbf{X}^T\mathbf{z}\mathbf{z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}/n] = \det(\mathbf{X}^T\mathbf{X})[1 - \mathbf{z}^T\mathbf{H}_c\mathbf{z}/n],$$

since $\det(\mathbf{I} - \mathbf{AB}) = \det(\mathbf{I} - \mathbf{BA})$ for any $\mathbf{A}$ and $\mathbf{B}$ with suitable dimensions. Hence, under the constraint $\mathbf{z} \in \mathbf{Z}$, solving the problem of $\min \det(\mathbf{W})$ is equivalent to solving the problem of $\max \mathbf{z}^T\mathbf{H}_c\mathbf{z}$. According to Theorem 2 in Li (2006), an optimal solution $\mathbf{z}$ to $\max_{\mathbf{z}\in\mathbf{Z}} \mathbf{z}^T\mathbf{H}_c\mathbf{z}$ is a sign eigenvector associated with the largest sign eigenvalue of $\mathbf{H}_c$. From Theorem 1, this partition is the same as the solution obtained by solving problem (4). This completes the proof.

A.3 Proof of Theorem 2

Differentiating the criterion function in (10) with respect to $\boldsymbol{\beta}$ and α, respectively, and equating derivatives to zero yields the least squares solution given in the theorem.

Next, we consider partition $\mathbf{z}$. Noting that $\mathbf{z}^T\mathbf{z} = n$, the numerator of the criterion function in (10) reduces to $(\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})^T(\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})-2\mathbf{z}^T(\alpha\mathbf{1}-\mathbf{X}\boldsymbol{\beta})+n$. Hence, maximising

the criterion with respective to $\mathbf{z}$ in (10) is equivalent to solving $\min_{\mathbf{z}\in\mathbf{Z}} \mathbf{z}^T(\alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta})$. Applying Theorem 2 in Li (2006) gives $\mathbf{z} = S(\alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta})$. Finally, by inserting the least squares solution $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}/(1-\lambda)$ and $\alpha = \overline{\mathbf{z}}$ into the above equation, we find that $\mathbf{z}$ is a sign eigenvector of $\mathbf{H}(1-\lambda)$ associated with the largest sign eigenvalue, satisfying $\mathbf{z} = S(\mathbf{H}(1-\lambda)\mathbf{z})$, where $\mathbf{H}(\tau) = \mathbf{H}_c + (\tau/n)\mathbf{1}\mathbf{1}^T$ and $\lambda = [\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]^T[\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}]/(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})$. This completes the proof.

A.4 Proof of Lemma 2

It is easy to show that vectors $\mathbf{1}$ and $-\mathbf{1}$ are ordinary eivenvectors of $\mathbf{H}(1)$ associated with the largest ordinary eigenvalue, 1. Hence they are sign eivenvectors of $\mathbf{H}(1)$ as well since they are sign vectors. To show that vectors $\mathbf{1}$ and $-\mathbf{1}$ are sign eigenvectors of $\mathbf{H}(1)$ associated with the *largest* sign eigenvalue, we note that any sign eigenvalue is not greater than the largest ordinary eigenvalue. This completes the proof.

References

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803-821.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C. , Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536-540.

Brusco, M. J. and Cradit, J. D. (2001) A variable selection heuristic for K-means clustering. *Psychometrika*, **66**, 249-270.

Celeux, G. and Govaert, G. (1993) Comparison of the mixture and classification maximum likelihood in cluster analysis. *Journal of statistical computation and simulation*, **47**, 127-146.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA.*, 95, 14863-14868.

Everitt, B. S. (1993) *Cluster Analysis.* New York: Arnold.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611-631.

Friedman, H. P and Rubin, J. (1967) On some invariant criteria for grouping data. *J. Amer. Statist. Assoc.*, **62**, 1159-1178.

Friedman, J. H. and Meulman, J. J. (2004) Clustering objects on subsets of attributes. *J. R. Statist. Soc.* B, **66**, 815-849.

Goldstein, D. R., Ghosh, D. and Conlon, E. (2002) Statistical issues in the clustering of gene expression data. *Statistica Sinica*, **12**, 219-240.

Hand, D. J. (2004) Discussion on 'Clustering objects on subsets of attributes' (by J. H. Friedman and J. J. Meulman). *J. R. Statist. Soc.* B, **66**, 839-840.

Hoff, P. D. (2004) Discussion on 'Clustering objects on subsets of attributes' (by J. H. Friedman and J. J. Meulman). *J. R. Statist. Soc.* B, **66**, 845-846.

Knoke, J. D. (1982) Discriminant analysis with discrete and continuous variables. *Biometrics*, **38**, 191-200.

Krzanowski, W. J. (1975) Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Assoc.*, **70**, 782-790.

Krzanowski, W. J. and Marriot, F. H. C. (1995) *Multivariate Analysis.* Part 2, New York: Arnold.

LaTulippe, E., Satagopan, J., Smith, A., Scher, H., Scardino, P., Reuter, V., and Gerald, W. L. (2002) Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Research*, **62**, 4499-4506.

Li, B. (2006) Sign eigenanalysis and its applications to optimizations and robust statistics. *Comput. Statist. Data Anal.*, **50**, 154-162.

MacQueen, J. B. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley: University of California Press, pp281-297.

SAS Institute (1999) SAS/STAT User's Guide. Cary, NC: SAS Institute Inc.

Satagopan, J. M. and Panageas, K. S.(2003) A statistical perspective on gene expression data analysis. *Statistics in Medicine*, **22**, 481-499.

Speed, T. (2003) *Statistical analysis of gene expression microarray data.* London: Chapman & Hall.

Ward, J. H. (1963) Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, **58**, 236-244.

Webb, A. (1999) *Statistical Pattern Recognition.* London: Arnold.

Weisberg, S. (1985) *Applied Linear Regression.* 2nd Ed. New York: Wiley.

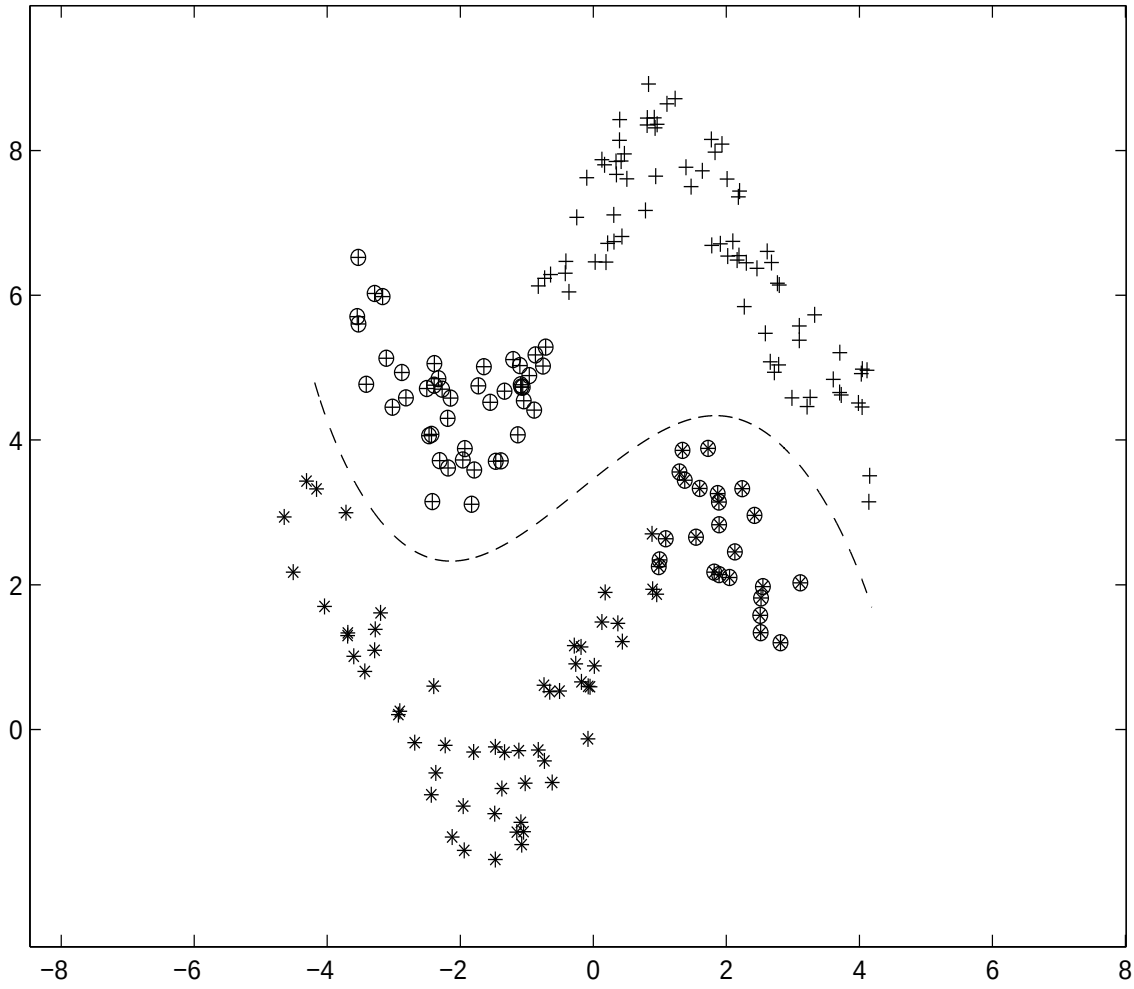

Fig. 1. Example 5.2: scatter plot of x_1 versus x_2 with a boundary (dashed line) of allocation regions by the clustering-function based method. Objects in group 1 denoted '*', objects in group 2 denoted '+'. The misclassified objects by the k-means method are circled.

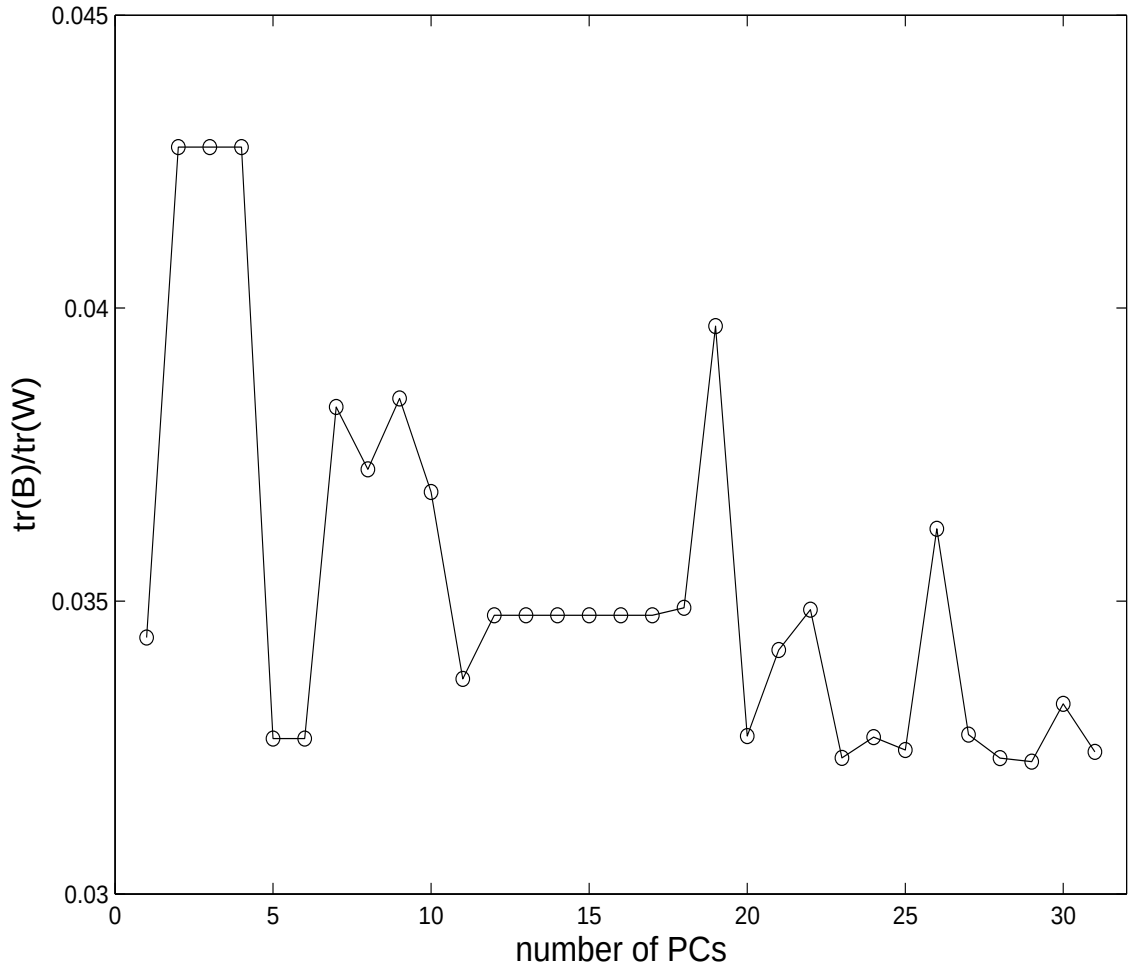

Fig. 2. trace($\mathbf{B}$)/trace($\mathbf{W}$) versus the number of principal components