Title:
Variable Selection for Clustering Analysis Based on Squared Euclidean Distance:
Analogy to Least Squares Regression

Author:
Hongmei JIANG

Author affiliations:
Department of Statistics, Northwestern University
2006 Sheridan Road, Evanston, IL 60208

Email address:
hongmei@northwestern.edu
Author’s Footnote

Hongmei Jiang is an Assistant Professor of Statistics (e-mail: hongmei@northwestern.edu),
Department of Statistics, Northwestern University, Evanston, IL 60208.
Abstract and Key Words

**ABSTRACT:** In this article we propose a variable selection method for non-model-based clustering. The basic idea is to conduct cluster analysis in the subspace of a subset of variables but to evaluate the partition of the samples in the original space. We define a criterion for variable selection based on squared Euclidean distance, which measures the proportion of variation in the data which can be explained by the clustering analysis based on a subset of variables. When the number of variables is large we propose a forward stepwise selection procedure to find a local optimum. We also propose a fast full search approach to find the global optimum or multiple sets of variables which define a data structure. The performance of the proposed method is demonstrated in both simulated and real data analyses.

**KEY WORDS:** Feature selection; Non-model-based clustering; Sum of squares; Unsupervised learning.


text

1 INTRODUCTION

Variable or feature selection is a challenging problem in both clustering and classification. The data structure may only be contained in a subset of available variables, while using all the variables for clustering or classification analysis may obscure the true structure. Fan & Fan (2008) demonstrated that classification using all features for high-dimensional data such as microarray gene expression data can be as poor as using random guessing, and it is very important to select the important variables for classification analysis. Further challenge in clustering is that the labels for the samples or objects are unobserved. Sometimes the number of clusters is unknown.

There are mainly two approaches for variable selection in clustering analysis in the literature. One approach performs variable selection and clustering analysis simultaneously, usually for model-based clustering. Raftery & Dean (2006) proposed a model-based clustering method to select variables, the number of clusters, and the clustering model simultaneously. Tadesse, Sha & Vannucci (2005) proposed a Bayesian variable selection method to uncover the cluster structure of the data and identify the discriminating variables for multivariate normal mixture model with unknown number of components. The second approach is to select a set of important discriminating variables first and then to perform clustering analysis on the selected variables only. Brusco & Cradit (2001) presented a variable-heuristic for K-means clustering based on the Adjusted Rand Index (Hubert & Arabie 1985) for measuring cluster recovery. Fraiman, Justel & Svarc (2008) proposed two procedures for variable selection in cluster analysis assuming that the number of cluster is known. Their
objective is to find a subset of variables which will reproduce the same partitions as
the multivariate partition procedure does using all the variables (100% efficiency).
Their method can also consider finding a subset of variables with possible reduction
in the efficiency to 95% or 90% of correct allocations of the samples. Eight differ-
ent variable selection techniques for model-based and non-model-based clustering are
compared in Steinley & Brusco (2008). Raftery & Dean (2006) also gave a list of
references on variable selection for both model-based and non-model-based clustering
in their discussion section.

In this article we introduce a variable selection method for non-model-based clus-
tering. The basic idea is to conduct cluster analysis using a subset of variables but
evaluate the partition or allocation of the samples or objects in the original space
with all available variables. Analogous to the coefficient of determination $R^2$ for the
least squares regression analysis, we define $G^2$ based on squared Euclidean distance,
which measures the proportion of variation in the data which can be explained by
the clustering analysis based on a subset of variables. The objective of the procedure
is to maximize $G^2$ with a smallest number of variables. Analogous to the adjusted
$R^2$ which penalizes the number of selected variables, we define the adjusted $G^2$ for
the clustering analysis. Using this criterion we seek to find a subset of variables for
which $G^2$ is at the maximum or close to maximum that adding more variables does
not help much on refining the clustering. Our proposed method can be used when
the number of clusters is known or unknown. For the latter case, we use the variance
ratio criterion due to Caliski & Harabasz (1974) to determine the optimal number
of clusters. When the number of variables is small, all the possible subsets of the
variables can be considered. When the number of variables is large we propose a
forward stepwise variable selection algorithm to find a local optimum.

Since our method is motivated by K-means clustering, a brief review on K-means clustering is given in Section 2.1. The criterion for variable selection is defined in Section 2.2. Variable selection procedures with known and unknown number of clusters are presented in Sections 2.3 and 2.4, respectively. These procedures include best subset selection, forward stepwise variable selection, and fast full search methods. Results comparing the performance of the proposed method with other methods are given in Section 3 for simulated data and Section 4 for several real data examples. Advantages and limitations of the method are discussed in Section 5.

2 METHODOLOGY

2.1 K-means Clustering

Let \((x_1, x_2, \cdots, x_n)\) denote the \(n\) observations or objects where each observation \(x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})\) is a \(p\)-dimensional real vector. The K-means clustering is a popular cluster analysis which partitions the \(n\) observations into \(K\) clusters or groups \((C_1, C_2, \cdots, C_K)\) where \(C_k\) is the set of \(n_k\) objects in cluster \(k\) \((k = 1, 2, \cdots, K)\), and \(K\) is pre-specified or given. It is applied in situations where all variables are of the quantitative type. The K-means method constructs the partitions by minimizing the squared Euclidean distance of the observations from the cluster means, that is minimizing the error sum of squares (SSE) or within-cluster sum of squares:

\[
\arg\min_{(C_1, C_2, \cdots, C_K)} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}^{(k)}\|^2 = \arg\min_{(C_1, C_2, \cdots, C_K)} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j^{(k)})^2, \tag{1}
\]
where $\bar{x}^{(k)}$ is the cluster mean for the $k$th cluster and is computed by averaging the values on each variable over the objects within the cluster:

$$
\bar{x}^{(k)} = (\bar{x}_1^{(k)}, \bar{x}_2^{(k)}, \ldots, \bar{x}_p^{(k)}) = \left( \frac{1}{n_k} \sum_{i \in C_k} x_{i1}, \frac{1}{n_k} \sum_{i \in C_k} x_{i2}, \ldots, \frac{1}{n_k} \sum_{i \in C_k} x_{ip} \right). 
$$

(2)

An iterative descent algorithm is typically used for K-means clustering. For an initial set of cluster means which could be randomly selected from the data set or produced by another clustering method, the K-means clustering algorithm proceeds by alternating two steps: assignment step and update step. In the assignment step, every observation is assigned to the cluster whose cluster mean is the closest. In the update step, new cluster means are calculated with the updated cluster membership. These two steps are iterated until the assignments of the observations to the clusters do not change. Since it is a heuristic algorithm the result may not represent a global optimum. It is very common to try several different sets of initial cluster means.

### 2.2 Criterion for Variable Selection

Since the criterion of K-means clustering is to minimize the error sum of squares, it is naturally desired to use the same criterion to select a subset of the variables for cluster analysis. When selection of variables with different dimensions is considered, not only the SSE needs to be minimized but also the variation of the data has to be taken into account. By analogy to the regression analysis, we can investigate the proportion of reduction of total variation associated with the clustering using the selected variables. Suppose a subset of variables are used for clustering, the objects are partitioned into different clusters. Then the variation of the data can be computed either for partial data such as the data associated with selected variables only, or for
the original complete data. Therefore we make a general definition of the proportion of reduction of variation in the data for the cluster analysis.

Let $T = \{1, 2, \ldots, p\}$ be the collection of the indexes of the $p$ variables, $S, V \subseteq T$ denote two subsets of the indexes. Let $X_i$ denote the $i$th variable where $i = 1, 2, \ldots, p$, and $X^S, X^V \subseteq \{X_1, X_2, \ldots, X_p\}$ denote the corresponding two sets of selected variables. For instance, when $S = \{1, 3\}$, $X^S = \{X_1, X_3\}$, the data associated are $x^S_i = (x_{i1}, x_{i3})$ for $i = 1, 2, \ldots, n$. Suppose a K-means clustering method is applied to group the subset of the data $x^S_i (i = 1, 2, \ldots, n)$ into $K$ clusters. Let $C^S(i)$ be the cluster assignment for the $i$th observation based on $X^S$. We would like to know what proportion of the variation in the data associated with $X^V$ (that is, $x^V_i (i = 1, 2, \ldots, n)$) can be explained by the clustering based on $X^S$. The measure of total variation, denoted by $\text{SST}$, is the sum of the squared distances of the observations from the grand mean of the data. Analogous to the coeffient of determination $R^2$ for the least squares regression analysis, we can define the proportion of variation in the data which can be explained by the clustering analysis.

**Definition 1.** The proportion of reduction of variation in the data associated with $X^V$ accounted by the clustering analysis based on the variables $X^S$ is:

$$G^2_K(S, V) = 1 - \frac{\text{SSE}_K(S, V)}{\text{SST}(V)} = 1 - \frac{\sum_{k=1}^{K} \sum_{j \in V} \sum_{C^S(i)=k} (x_{ij} - \bar{x}^S_{kj})^2}{\sum_{j \in V} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$

(3)

where $K$ is the number of clusters, $\bar{x}^S_{kj} = \sum_{C^S(i)=k} x_{ij} / n^S_k$ is the cluster mean for the $k$th cluster ($k = 1, 2, \cdots, K$) and the $j$th variable ($j \in V$), $n^S_k$ is the number of objects in cluster $k$ for the clustering based on $X^S$, and $\bar{x}_j = \sum_{i=1}^{n} x_{ij} / n$ is the grandmean for the $j$th variable ($j \in V$).
Notice that $0 \leq G^2 \leq 1$ and minimizing SSE is equivalent to maximizing $G^2$.

There are two interesting special cases for $G^2$. When $V = S$, we have,

$$G^2_K(S, S) = 1 - \frac{SSE_K(S, S)}{SST(S)} = 1 - \frac{\sum_{k=1}^{K} \sum_{j \in S} \sum_{C^S(i) = k} (x_{ij} - \bar{x}_j)^2}{\sum_{j \in S} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}. \quad (4)$$

Then $G^2_K(S, S)$ emphasizes that both clustering and computation of the variation use the subset of the data associated with the selected variables $X^S$ only. When $V = T$, we have

$$G^2_K(S, T) = 1 - \frac{SSE_K(S, T)}{SST(T)} = 1 - \frac{\sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{C^S(i) = k} (x_{ij} - \bar{x}_j)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}. \quad (5)$$

In this case, partial data associated with $X^S$ are used for cluster analysis and the membership assignment is obtained for each of the $n$ observations. Within each cluster the cluster mean can be computed for each of the $p$ variables, then the variation of the data is computed using the original complete data.

The properties and the performances of $G^2$ are investigated for the above two special cases using a simple example from Brusco & Cradit (2001) which has 9 objects measured on 4 variables (Table 1). Since variables $X_1$ and $X_2$ define a clear and fine structure for the data we will focus on these two variables for the moment (Figure 1). Using both $X_1$ and $X_2$, $X_1$ only, $X_2$ only, the data points are partitioned into three clusters (Figure 2), and the corresponding $G^2$ (subscript 3 is omitted for this example) are computed below:

$$G^2(\{1, 2\}, \{1, 2\}) = 0.9713,$$
$$G^2(\{1\}, \{1\}) = 0.9607, \quad G^2(\{1\}, \{1, 2\}) = 0.9713,$$
$$G^2(\{2\}, \{2\}) = 0.9877, \quad G^2(\{2\}, \{1, 2\}) = 0.6331.$$
This example sheds some lights on choosing the best criterion for variable selection. Obviously, finding a subset of $S$ with a smallest number of cardinality to maximize $G^2(S, S)$ does not work. It can be seen that $G^2(\{2\}, \{2\})$ gives the largest $G^2$ value however it makes a wrong assignment of the clusters — observations 4 and 9 are put into wrong clusters (Figure 2 (g) and (h)). Another reason that $G^2(S, S)$ is not a good choice is that it increases when the dimension of $S$ increases even when the additional variable does not contribute more information in clustering and the clusters do not change. In this example, using $X_1$ only can do a perfect job to partition the objects. When we use both $X_1$ and $X_2$, the clustering results are exactly the same as those based on $X_1$ only, however $G^2(\{1, 2\}, \{1, 2\})$ is greater than $G^2(\{1\}, \{1\})$. But if we use $G^2(S, T) = G^2(S, \{1, 2\})$ as the criterion, we can see that $G^2(\{2\}, \{1, 2\})$ is smallest, and $G^2(\{1\}, \{1, 2\}) = G^2(\{1, 2\}, \{1, 2\})$ because both $X_1$ and $X_2$, and $X_1$ only produce the same clustering results. Therefore we propose to select the smallest subset of variables $S$ such that $G^2_K(S, T)$ is maximized:

$$
\arg\max_{S \subseteq \{1, 2, \ldots, p\}} G^2_K(S, T) = \arg\max_{S \subseteq \{1, 2, \ldots, p\}} 1 - \frac{\sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{C^S(i)=k} (x_{ij} - \bar{x}_{kj})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})^2}.
$$

(6)

Since the denominator SST is constant for all possible subsets of variables, $G^2$ varies inversely with SSE. Therefore this is equivalent to find the smallest set of $S$ such that

$$
SSE_K(S, T) = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{C^S(i)=k} (x_{ij} - \bar{x}_{kj}^S)^2
$$

is minimized.
Definition 2. Analogous to the adjusted $R^2$ which penalizes the number of selected variables, we define the adjusted $G^2$ for the clustering analysis as:

$$G^2_{K,a}(S, T) = 1 - \frac{\text{SSE}_K(S, T)}{\text{SST}(T)}/(n - |S|) = 1 - \frac{\sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{C^S(i)=k} (x_{ij} - \bar{x}_{kj})^2/(n - |S|)}{\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})^2/(n - 1)}, \quad (7)$$

where $|S|$ is the number of indexes in $S$, that is, the number of variables selected for cluster analysis, $T$ is the collection of all the available variables. Other notations are the same as in Definition 1.

Using this criterion we seek to find a subset of variables for which $G^2$ is at the maximum or close to maximum that adding more variables does not help much on refining the clustering.

2.3 Variable Selection with Known Number of Clusters

We first assume that the number of clusters, $K$, is known.

Proposition 1. If the clustering analysis based on two different sets of variables, $S_1$ and $S_2$, gives the same clustering assignments, then $G^2_K(S_1, T) = G^2_K(S_2, T)$.

Proof. Since the two sets of variables give the same clustering assignment, the membership for the $i$th observation is the same for the two separate clustering, that is, $C^{S_1}(i) = C^{S_2}(i)$ for each $i = 1, 2, \cdots, n$. The cluster means for each of the $K$ clusters are also the same, that is, $\bar{x}_{kj}^{S_1} = \bar{x}_{kj}^{S_2}$ for $j = 1, 2, \cdots, p$ and $k = 1, 2, \cdots, K$. 

11
Therefore,

\[ G^2_{K}(S_1, T) = 1 - \frac{\sum_{k=1}^{K} p \sum_{j=1}^{p} \sum_{i=1}^{C^{S_1}(i)=k} (x_{ij} - \bar{x}^{S_1}_{kj})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \]

\[ = 1 - \frac{\sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{C^{S_2}(i)=k} (x_{ij} - \bar{x}^{S_2}_{kj})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \]

\[ = G^2_{K}(S_2, T). \]

\( \square \)

**Proposition 2.** If K-means method is used for the clustering analysis, then \( G^2_{K}(T, T) = \max_{S \subseteq T} G^2_{K}(S, T). \)

**Proof.** Since the purpose of the K-means clustering is to construct the partitions by minimizing the SSE or within-cluster sum of squares, the clustering using all the \( p \) variables will give the smallest SSE when the complete data are considered, hence the largest \( G^2 \). When a subset of variables \( S \) is used for K-means clustering, the clustering results in the smallest SSE for the partial data which are associated with \( S \). If it gives the same clustering assignments as using \( T \), then \( G^2_{K}(T, T) = G^2_{K}(S, T) \) according to Proposition 1. On the other hand, if \( S \) gives different clustering assignments from using \( T \), then it will give a larger SSE than \( T \) when the complete data are considered.

Therefore \( G^2_{K}(T, T) = \max_{S \subseteq T} G^2_{K}(S, T). \) \( \square \)

**Theorem 1.** When maximizing \( G^2_{K}(S, T) \) is used as the criterion for variable selection for K-means clustering, it is equivalent to finding a subset of variables \( S \) which produces the same clustering assignments as using all the variables.
The proof is obvious by combining the results of Propositions 1 and 2.

2.3.1 Best subset selection

When \( p \) is small, all the possible subsets of the variables can be considered. Then the variable selection procedure is to find one best subset of variables with smallest number of variables to maximize \( G^2_K(S, T) \). When all the possible combinations of variables are considered we call the variable selection as best subset selection method.

2.3.2 Forward stepwise variable selection

When the number of variables \( p \) is large we would not be able to enumerate all the possible subsets of the variables. For example when \( p = 20 \), there are more than one million \((2^{20} - 1)\) possible subsets based on the assumption that each variable is either included or excluded. Here we propose a forward stepwise variable selection algorithm.

1. Initial step: First the K-means clustering algorithm is applied for each of the \( p \) variables separately and \( G^2_K(j, T) \) is computed for each \( j = 1, 2, \cdots, p \). The \( X \) variable with the largest \( G^2 \) is then selected. Suppose \( G^2_K(j_1, T) \) is the largest among all the \( G^2_K(j, T) \) \((j = 1, 2, \cdots, p)\) then let \( S = \{j_1\} \).

2. Addition step: Next we will consider which variable should be added to \( S \). Perform cluster analysis with two \( X \) variables, where \( X_{j_1} \) is one of the pair. Then find \( j_2 \) such that

\[
G^2_K(S \cup j_2, T) = \max_{j \in T, j \notin S} G^2_K(S \cup j, T).
\]
Then $j_2$ will be added to $S$ if

$$G^2_K(S \cup j_2, T) > G^2_K(S, T);$$

otherwise $j_2$ will not be added to $S$. Here $S \cup j_2$ is the union of $S$ and $j_2$.

3. Removal step: After a new variable is added we will examine if any of the other variables should be dropped from the selection. For each variable $X_j \in X^S$, cluster analysis is performed using all the variables in $X^S$ except for $X_j$. Then $G^2_K(S \setminus j, T)$ computed for each $j \in S$, where $S \setminus j$ means all the variables in $S$ except for $j$. Variable $X_j$ will be dropped if $G^2_K(S \setminus j, T) \geq G^2_K(S, T)$; otherwise $X_j$ is retained.

4. Steps 2 and 3 will be iterated for examining which variable is the next candidate for addition, then examining which variables already selected should be dropped. This search process will stop until no more variables can either be added or dropped. Then $S$ is the set of selected variables.

2.3.3 Fast full search

Using the forward stepwise selection procedure one might get a set of variables $S$ with the corresponding $G^2$ as the local maximum but not the global maximum. This phenomena is demonstrated in the data analysis examples 1 and 2 (Section 4). One reason is that the variables are selected one by one in the procedure. After several steps into the algorithm, adding any one of the remaining variable may not increase the $G^2$ value. In this case the search is stopped. However adding one variable alone does not improve the clustering does not mean that this variable is not important. When several variables which do not have big contribution in clustering by themselves
work together a better clustering result may be obtained. Another disadvantage of forward stepwise selection method is that it only produces one set of variables. In practice there may exist more than one set of variables which define the same data structure, such as the microarray data example in this article. To reach the global maximum and/or get more than one set of variables, we propose a fast full search. The main idea is to use each of the $p$ variables as the initial starting point and to obtain a corresponding set of selected variables by applying the forward stepwise selection algorithm. Then $p$ sets of variables will be produced, and the set or sets of variables which give the largest $G^2$ will be the final result. The following give the details.

1. For $j = 1$, the initial starting set contains the first variable only. Then apply Steps 2-4 of the forward stepwise variable selection algorithm and obtain a set of selected variables denoted as $S_1$. In other words, the first variable entering the set is not selected by the univariate analysis but fixed.

2. For variable $j = 2, 3, \cdots, p$, obtain $S_j$ as in the first step. That is, for variable $j$, the initial starting set contains the $j$th variable only. Then apply Steps 2-4 of the forward stepwise variable selection algorithm and obtain a set of selected variables denoted as $S_j$.

3. The set or sets of the selected variables are then $\arg \max_{S_j \in \{1, 2, \cdots, p\}} G^2(S_j, T)$.

One can also compute how frequently each of the $p$ variable is selected in the $p$ searches. This frequency indicates how important a variable is in defining the data cluster structure. It might be tempting to use the variables with high frequencies as the selected variables. However this involves how to decide the cut-off points. If one
does not want to miss any important variable, then one can simplify use the union of the \( p \) sets of the variables. That is, only the variables which have never been selected in the \( p \) searches will not enter the set of the selected variables.

In the above best subset selection and forward stepwise selection procedures, we use \( G^2 \) as the selection criterion. When comparing two subsets of variables \( S_1 \) and \( S_2 \), \( S_1 \) is preferred when \( G^2_K(S_1,T) > G^2_K(S_2,T) \), or when \( G^2_K(S_1,T) = G^2_K(S_2,T) \) and \( S_1 \) contains smaller number of variables. Here the complete data set is used as the reference. Even though this is the method used in all the simulation studies and real data analyses in this paper, it might also be possible to use the data combining the two subsets of variables as an alternative reference. To be specific, \( S_1 \) is preferred when \( G^2_K(S_1, S_1 \cup S_2) > G^2_K(S_2, S_1 \cup S_2) \), or when \( G^2_K(S_1, S_1 \cup S_2) = G^2_K(S_2, S_1 \cup S_2) \) and \( S_1 \) contains smaller number of variables.

Sometimes a subset of variables \( S \) gives a \( G^2 \) which is not equal to but close to the maximum, and adding more variables does not help much on improving the clustering. If smaller number of variables is preferred, we may select \( S \) even though the corresponding \( G^2 \) is not at the maximum. In this situation \( G^2_a \) can be used as the criterion for variable selection as it takes into account the number of selected variables. For this purpose both the best subset selection and forward stepwise selection procedures can be used when the criterion of \( G^2 \) is replaced by \( G^2_a \).

### 2.4 Variable Selection with Unknown Number of Clusters

One challenging problem in unsupervised clustering is the unknown number of clusters. In this section we develop a method to combine the cluster number determination and variable selection for clustering analysis. Different methods have been developed
to estimate the optimal number of clusters. About 30 procedures were compared and
examined in Milligran & Cooper (1985), one recommended method is the variance
criterion due to Caliski & Harabasz (1974):

\[ VRC_K = \frac{BGSS}{K-1} \cdot \frac{WGSS}{n-K}, \]

where BGSS and WGSS are the between- and within-group (or cluster) sums of
squares, with \( K \) clusters and \( n \) observations. Using our notation in the paper, WGSS
= SSE, and BGSS = SST-SSE. Using this criterion \( K \) is chosen to maximize \( VRC_K \).
As Tibshirani, Guenther & Trevor (2001) pointed out that this ratio is not defined
for \( K = 1 \). We use this criterion by assuming that there are at least 2 clusters in the
data set.

For the best subset selection method, we examine all the possible combinations
of the \( p \) variables. Realizing that the number of clusters depend on the selected
variables, for each subset \( S \) we first determine the optimal number of clusters using
the variance ratio criterion. To be specific, K-means clustering is performed using the
set of variables \( S \) for \( K = 2, 3, \ldots, G_{max} \) clusters respectively, where \( G_{max} \) is a fixed
number. Then the variance ratio criterion is computed as:

\[ VRC_K(S) = \frac{SST(S) - SSE_K(S, S)}{K-1} / \frac{SSE_K(S, S)}{n-K}. \]  

Notations of \( SST(S) \) and \( SSE_K(S, S) \) are used to emphasize that both the clustering
and computation of variation is limited to the partial data associated with \( S \) only.
Then the optimal number of cluster \( K_S \) is chosen such that \( VRC_K(S) \) reaches a local
maximum if not global maximum at \( K_S \). For each subset \( S \) and the chosen \( K_S \) and
the corresponding clustering assignments, \( G_{K_S}^2(S, T) \) can be computed. The best
subset of variable \( S \) is the one at which \( G_{K_S}^2(S, T) \) is maximized.
For forward stepwise selection method, we need add one more step about the cluster size determination.

1. Initial step: First the clustering analysis is performed for each of the $p$ variables for $K = 2, 3, \ldots, G_{\text{max}}$ clusters respectively, and $VRC_K(j)$ is computed for $j = 1, 2, \ldots, p$. For each $j$, the optimal number of clusters $K_j$ is chosen such that $VRC_K(j)$ reaches the first local maximum over the range of $K$. Then $G^2_{K_j}(j, T)$ is computed for each $j = 1, 2, \ldots, p$. The $X$ variable with the largest $G^2$ is then selected. Suppose $G^2_{K_{j_1}}(j_1, T)$ is the largest among all the $G^2_{K_j}(j, T)$ $(j = 1, 2, \ldots, p)$ then let $S = \{j_1\}$ and the number of clusters $K_S = K_{j_1}$.

2. Addition step: Next we will consider which variable should be added to $S$. Perform cluster analysis with two $X$ variables, where $X_{j_1}$ is one of the pair. For each pair of variables, the optimal number of clusters is determined by maximizing $VRC_K(S \cup j)$ over the range of $K = 2, 3, \ldots, G_{\text{max}}$ for $j \in T \setminus S$. Then find $j_2$ such that

$$G^2_{K_{S \cup j_2}}(S \cup j_2, T) = \max_{j \in T, j \notin S} G^2_{K_{S \cup j}}(S \cup j, T).$$

Then $j_2$ will be added to $S$ and the number of cluster is updated to $K_{S \cup j_2}$ if

$$G^2_{K_{S \cup j_2}}(S \cup j_2, T) > G^2_{K_S}(S, T);$$

otherwise $j_2$ will not be added to $S$ and $K_S$ does not change.

3. Removal step: After a new variable is added we will examine if any of the other variables should be dropped from the selection. For each variable in $X_j \in X^S$, cluster analysis is performed using all the variables in $X^S$ except for $X_j$ and the optimal number of cluster is determined by maximizing $VRC$.
Then $G^2_{K_S \setminus j}(S \setminus j, T)$ computed for each $j \in S$. Variable $X_j$ will be dropped if $G^2_{K_S \setminus j}(S \setminus j, T) \geq G^2_{K_S}(S, T)$, and $K_S$ is updated to $K_{S \setminus j}$; otherwise variable $X_j$ is retained and the number of clusters $K_S$ does not change.

4. Steps 2 and 3 will be iterated for examining which variable is the next candidate for addition, then examining which variables already selected should be dropped. This search process will stop until no more variables can either be added or dropped. Then $S$ is the set of selected variables and $K_S$ is the optimal number of clusters.

3 Simulation

3.1 Simulation 1: small $p$

Framan et al. (2008) did a Monte Carlo study for K-means clustering of simulated data with three variables. They also compared their results with other variable selection methods proposed by Brusco & Cradit (2001), Raftery & Dean (2006), and Steinley & Brusco (2008). We use the same set up as in the simulation study of Framan et al. (2008) so that we can compare our results with the results presented in their paper (especially their Tables 1 and 2). We generate the observations in a three-dimensional space with a mixture of normal distribution,

$$X = (X_1, X_2, X_3) \sim \sum_{i=1}^{3} \alpha_i N_3(\mu_i, \Sigma_i),$$

where $\alpha_1 = \alpha_2 = 0.35$, and $\alpha_3 = 0.30$, and $X_1$ and $X_2$ are independently generated in the following way:

$$X_1 \sim \alpha_1 N(0, 0.2) + \alpha_2 N(0.1, 0.2) + \alpha_3 N(0.9, 0.2),$$
\[ X_2 \sim \alpha_1 N(0, 0.2) + \alpha_2 N(0.9, 0.2) + \alpha_3 N(0.1, 0.2). \]

For \( X_3 \), two cases are considered.

Case 1: \( X_3 \) is an independent “noisy” variable: \( X_3 \sim N(0, \sigma), \sigma = 0.1, 0.2, 0.3. \)

Case 2: \( X_3 \) is linearly correlated with \( X_1 \) and \( X_2 \): \( X_3 = (X_1 + X_2) / \sqrt{2}. \)

In each situation 1000 simulated data sets each with size \( n=100 \) are generated. Since the number of variables, 3, is very small we can use the best subset selection method. For the purpose of demonstration we also use the stepwise forward selection procedure. In fact these two methods give exactly the same result. We also compute the “misclassification” error rate by comparing the clustering assignments with the true underlying distribution (equation [10]). We want to point out here that outliers are not removed for any of the simulated data. If a data point is generated under the first component of the mixture model but closer to the second component as an outlying observation, it will be assigned to the cluster representing the second component and counted as a misclassification error.

We first perform the K-means clustering assuming that the number of clusters \( K \) is known as 3. Table [2] shows how frequently 2 variables or 3 variables are selected. For Case 1 where \( X_3 \) is an independent random noise, \( X_1 \) and \( X_2 \) are selected 100% of the times whenever 2 variables are selected. When the variation of \( X_3 \) increases, there is an increasing pattern for selecting all 3 variables. when \( G^2 \) criterion is used for variable selection, the optimal solution of selecting \( X_1 \) and \( X_2 \) is achieved in 96% of the replicates for \( \sigma = 0.1 \) and in 72% of the replicates for \( \sigma = 0.3 \). While when the adjusted \( G_{a}^{2} \) criterion is used for variable selection, the optimal solution of selecting \( X_1 \) and \( X_2 \) is achieved in 100% of the replicates for both \( \sigma = 0.1 \) and \( \sigma = 0.2 \), and in 99% of the replicates for \( \sigma = 0.3 \). It also tells us that the \( G^2 \) is very close for
the two sets of variables \( \{X_1, X_2\} \) and \( \{X_1, X_2, X_3\} \). When the number of variables is taken into account as in \( G^2_a \), \( \{X_1, X_2\} \) is preferred than \( \{X_1, X_2, X_3\} \). We also compare our results with Table 1 of Fraiman et al. (2008). It is clear that our method outperforms the methods of Brusco & Cradit (2001) and Steinley & Brusco (2008), and is comparable with the method of Fraiman et al. (2008). For Case 2 where multicollinearity exists, there is more chance to select 3 variables. To be specific, 2 variables are selected in 38% of the replicates when \( G^2 \) is used, and in 69% of the replicates when \( G^2_a \) is used. This tells us that \( X_3 \) also provides some information for clustering analysis. It performs better than the method of Brusco & Cradit (2001) which selects 3 variables in 100% of the replicates, and Steinley & Brusco (2008) which selects 2 variables more often but also tends to select 1 variable only and yields higher misclassification rate (Table 1 of Fraiman et al. (2008)). Our method performs better than the method of Fraiman et al. (2008) when 100% efficiency is used, but a little bit worse than the latter when 95% or 90% efficiency is used.

For the simulated data the second clustering analysis is done without the assumption that the number of clusters is known, by performing number of clusters determination and variable selection simultaneously. During the process, the variance ratio criterion is used to determine the optimal cluster size over the range of \( \{2, 3, \cdots, 8\} \) (i.e., \( G_{max} = 8 \)), the adjusted \( G^2_a \) is used for variable selection. The results from forward variable selection procedure are presented in Table 3. For all the situations, the number of clusters is dominately chosen as 3. This simulation verifies that that the variance ratio criterion has a very good performance in determining the
optimal number of clusters. Comparing with the method of Raftery & Dean (2006) for which the results are presented in Table 2 of Fraiman et al. (2008) for the same simulation set up, our method performs better for $\sigma = 0.2$ of Case I and Case II. 

For all the simulations and data analyses in the paper, function `kmeans` in R package is used to perform K-means clustering with the default algorithm of Hartigan & Wong (1979) and 25 randomly chosen starts.

[Table 3 is about here].

### 3.2 Simulation 2: large $p$

We also examine the performance of the proposed method when $p$ is relatively large. Based on Simulation 1, we generate 97 more independent noisy variables from $N(0, \sigma), \sigma = 0.1, 0.2, 0.3$ for both Case 1 and Case 2. Then in total there are $p = 100$ variables. In each situation 100 simulated data sets each with size $n = 100$ are generated. When the number of clusters is assumed to known as $K = 3$, three methods are applied to the clustering analysis: 3-means clustering using all variables, the proposed forward stepwise selection method using $G^2$ as the criterion, and the proposed forward stepwise selection method using $G^2_a$ as the criterion. When the number of clusters is assumed to be unknown, the proposed forward stepwise selection method with unknown number of clusters (Section 2.4) is used. The results are presented in Tables 4 and 5. In general, when $\sigma$ increases the misclassification rate increases for all four presented methods. However when all the variables are used for clustering the misclassification error increases dramatically, from 1.9% for $\sigma = 0.1$ to 26% for $\sigma = 0.3$. This demonstrates that the true data structure can be obscured with the accumulation of the random noise when all the variables are used for clustering. For
fixed number of clusters $K = 3$, the proposed forward stepwise selection method has smaller misclassification than that using all the variables. Especially for $\sigma = 0.3$ of Case 1, the proposed forward stepwise selection method with $G^2$ as the criterion has a misclassification error rate of 3%, while clustering using all variables has a misclassification error rate of 26%. When the number of clusters is assumed to be unknown, the variance ratio criterion is not able to determine the optimal number of clusters for the clustering analysis using all $p = 100$ variables. However the variance ratio criterion works well in the proposed forward stepwise selection method with unknown number of clusters. The proposed method finds the correct number of clusters and produces small misclassification error for both $\sigma = 0.1$ and 0.2. But when the variation of the random noisy variables is larger than the discriminating variables $X_1$ and $X_2$, that is, when $\sigma = 0.3$, the variance ratio criterion does not give the correct number of clusters which in turn affects the performance of the variable selection method.

4 Examples

4.1 Brusco and Cradit Data

We revisit the Brusco and Cradit data (Table 1) and this time we use all four variables. The pairs plot of these four variables is given in Figure 1. As we have seen before variables $X_1$ and $X_2$ define a cluster structure (cluster 1 contains observations 1, 2, and 3, cluster 2 contains observations 4, 5, and 6, and cluster 3 contains observations 7, 8, and 9) and we can use $X_1$ to assign the observations into the corresponding clusters correctly. Table 6 shows the $G^2$ and $G^2_a$ values for all the possible combination of the subsets. When the criterion $G^2$ is considered, using all four variables gives the largest
\( G^2 \) which is 0.6076, however observation 7 is assigned to the cluster which consists of observations 4, 5, and 6. The next largest value of 0.5890 is produced by using one variable \( X_1 \) only, or two variables \( X_1 \) and \( X_2 \), or three variables \( X_1, X_2, \) and \( X_3 \), or \( X_1, X_2, \) and \( X_4 \), and all these subsets of variables correctly assigned the observations into the corresponding clusters defined by the variables \( X_1 \) and \( X_2 \). When \( G^2_a \) is used as the criterion, \( X_1 \) only is the selected variable for the cluster analysis.

Table 6 is about here.

The forward stepwise selection method is also applied in this simple example. No matter what criterion \( G^2 \) or \( G^2_a \) is used, variable \( X_1 \) is always selected in the first step. For the next addition step, since no variable will contribute to the cluster analysis further (note \( X_1 \) and \( X_2 \) together give the same partition as using \( X_1 \) only), the search process will stop and the selected variable is \( X_1 \) only. In this example, the four variables together define a different data structure than that by using three or less variables. Our proposed method performs well intuitively and gives the correct discriminating variable for the data clusters presented in a subspace.

### 4.2 Wine Data

The wine data (Asuncion & Newman 2007) contain the results of a chemical analysis of three types of wines grown in the same region in Italy but derived from different cultivars. There are 13 measurements for each of the 178 samples of wines and they are: 1) Alcohol, 2) Malic acid, 3) Ash, 4) Alcalinity of ash, 5) Magnesium, 6) Total phenols, 7) Flavanoids, 8) Nonflavanoid phenols, 9) Proanthocyanins, 10) Color intensity, 11) Hue, 12) OD280/OD315 of diluted wines, and 13) Proline. Each of the
13 variables is standardized by dividing the corresponding sample standard deviation as in Dy & Brodley (2004). When all 13 variables are used for clustering analysis with 3 clusters, 6 samples of the second type of wine is clustered with the first and third type of wines with 3 each. The confusion matrix is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>59</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Type 2</td>
<td>3</td>
<td>65</td>
<td>3</td>
</tr>
<tr>
<td>Type 3</td>
<td>0</td>
<td>0</td>
<td>48</td>
</tr>
</tbody>
</table>

Since there are 8191 possible combinations of the 13 variables, we perform variable selection using the forward stepwise selection procedure instead of the best subset selection procedure. With a fixed number of clusters \( K = 3 \), variables 7, 1, 11 and 13 are selected sequentially for both \( G^2 \) and \( G^2_0 \) criterion and the confusion matrix is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>57</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Type 2</td>
<td>1</td>
<td>66</td>
<td>4</td>
</tr>
<tr>
<td>Type 3</td>
<td>0</td>
<td>0</td>
<td>48</td>
</tr>
</tbody>
</table>

With a reduction of variable dimension from 13 to 4, the misclassification error rate only increases by 1/178. When both the number of clusters determination and variable selection are performed simultaneously, the optimal number of cluster is chosen as 3 with \( G_{max} = 6 \) and the selected variables are also 7, 1, 11 and 13. Using these four variables the \( G^2 \) value is 0.4417.

The fast full search is also applied using each variable as the starting point. With 13 variables 13 sets of selected variables are obtained. The following table shows how
frequently a variable is selected in the 13 searches. It can be seen that variables 7, 1, 11 and 13 selected by our proposed forward stepwise selection method have high frequencies being visited in the fast full search, and variables 3, 4, and 8 are not selected for any of the 13 sets. The search starting with variable 10 gives the largest \( G^2 \) as 0.4477, and selects variables 10, 7, 13, 12, 11, 5, 2, 1, 6, 9, sequentially. It is just a coincidence that the union of the 13 sets of variables is the same as the search using variable 10 as the starting point. With a reduction of variable dimension from 13 to 10, the same clustering partitions are produced as using all 13 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>13</td>
<td>6</td>
<td>13</td>
</tr>
</tbody>
</table>

### 4.3 Microarray Data

The breast cancer data by van’t Veer & et al. (2002) contain gene expression measurements on 24,481 genes for 98 patients: 34 of them developed distant metastases within 5 years, 44 of them continued to be disease-free after a period of at least 5 years, 18 from BRCA1 mutation carriers, and 2 from BRCA2 carriers. Using unsupervised hierarchical clustering on approximately 5000 genes which were statistically significantly regulated across the groups of samples, the authors clustered the 98 patients into 2 groups which are associated with oestrogen receptor (ER)-\( \alpha \) status.

When 2-means clustering is performed using all 24,481 genes there are 55 (27 ER negative and 28 ER positive) and 43 (12 ER negative and 31 ER positive) samples in each of the two clusters. In fact many out of the 24,481 genes are not correlated with ER status and their inclusion in the clustering analysis makes it harder to identify
the true data structure. Fan & Fan (2008) demonstrated similar phenomena in high-dimensional classification where the classification using all features can be as poor as using random guessing. Therefore our first step is to use a filtering approach to reduce the number of genes. Following many other authors, such as Krzanowski & Hand (2009), we assume that if a cluster structure is present for a set of genes then it also appears on the individual gene level. Based on the significance of the clustering analysis on individual genes (see Appendix), we filter through $p = 158$ candidate genes, which define a better clustering structure for ER status.

Applying 2-means clustering using these 157 genes we get 2 clusters with sizes of 96 and 2 respectively, where samples 14 and 15 are the only samples in the second cluster. With 3-means clustering, samples 14 and 15 are still the only samples in one cluster, and the confusion matrix is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER negative</td>
<td>31</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>ER positive</td>
<td>1</td>
<td>56</td>
<td>2</td>
</tr>
</tbody>
</table>

The clustering structure based on $p = 158$ filtered genes is therefore better reflecting the ER status. We can further reduce the number of genes by applying the forward stepwise variable selection and fast full search approaches described in this paper. Two subsets of genes (set 1: CTSC, HNF3A, NM_005310, and ESR1; set 2: Contig46934RC, HNF3A, CEGP1 and FLJ20435) with only one gene (HNF3A) in common are identified, and each of the two sets of genes produce the same clustering results as using all $p = 158$ genes. This indicates that for a high-dimensional data there may exist more than one set of genes which define the same cluster structure. In either case, our proposed method has successfully reduced the number of candidate
genes from $p = 158$ to 4, without affecting the good clustering performance in relation to ER status.

5 Discussions

We have proposed a variable selection method for non-model-based clustering. The main idea is to conduct cluster analysis using a subset of variables but evaluate its performance in the original space using all available variables. When the number of variables is large we propose a forward stepwise selection procedure to find a local optimum. We also propose a fast full search approach to find the global optimum or more than one sets of variables which define the data structure. We demonstrate in numerical examples that our method successfully reduces the number of clustering variables without compromising the clustering performance.

When $G^2$ is used as the criterion our method is similar to Fraiman et al. (2008), because the purposes of the two methods are the same: finding a subset of variables which have the same clustering assignment as using all variables. However, the current work is different in several ways. First of all, similar to $R^2$ for least squares regression, the definition $G^2$ itself is a measurement on how good the clustering assignment is. While Fraiman et al. (2008) uses the clustering result with no variable selection as the reference, but it does not have an absolute measurement of how close the observations within each cluster are. Second, our method can combine both variable selection and estimation of the number of clusters simultaneously. Third, different ways of penalizing the number of selected variables are used. We use adjusted $G^2_a$ which takes into account the number of variables in a natural way. Fraiman et al. (2008)
use reduced efficiency rate, but they do not specify which rate to choose for a given situation.

In all the simulation studies and the real data analyses presented in this paper, K-means clustering is used. But the same principle can be applied to other non-model-based or model-based clustering methods, and evaluations of the proposed method are in progress. When the number of clusters is unknown, we use the variance ratio criterion to estimate the optimal number of clusters due to its simple calculation. Other methods, such as the Gap Statistic (Tibshirani et al. 2001) can also be used. In Simulation 2 with $p = 100$ and $n = 100$, we find that the variance ratio criterion can not determine the number of clusters when all variables are used. It also gives less accurate estimation when the variation of the noisy variables is larger than the discriminating variables. Whether other methods can give an accurate estimate of the number of clusters in these situations needs further study.
Appendix: Filtering Approach for Microarray Data Analysis

We assume that gene expression levels are normally distributed and perform the standardization for each gene by subtracting the corresponding mean and dividing the standard deviation. Let $x_{ij}$ denote the gene expression for the $i$th sample ($i = 1, 2, \ldots, n$) and the $j$th gene ($j = 1, 2, \ldots, p$). The standardized gene expression is denoted by $y_{ij} = (x_{ij} - \bar{x}_j)/s_j$, where $\bar{x}_j$ and $s_j$ are the sample mean and sample standard deviation for the $j$th gene. Then 2-means clustering is performed for each of the $p$ genes and $G^2_2(j, j)$ is computed for $j = 1, 2, \ldots, p$. Based on these $G^2_2(j, j)$ values we would like to select a subset of genes. To do this we need to know the distribution of $G^2_2(j, j)$ under the null hypothesis that the gene expression has a normal distribution. We randomly generated gene expression levels $z_{il}$ from $N(0, 1)$ with the same sample size $n$ for variable $l = 1, 2, \ldots, m$, where $m$ is a large number and chosen as $10^5$ in our analysis. Then 2-means clustering is performed on $z_{il}$ for each $l$ separately and $G^2_2(l, l)$ is computed. We then compute $p_j = \#\{l : G^2_2(l, l) \geq G^2_2(j, j)\}/m$ for each $j$. If the $j$th gene does not contribute in defining the cluster structure we expect $p_j$ to be relatively large; otherwise $p_j$ is small. Bonferroni correction method is applied to these $p_j$ values and 158 genes are identified as statistically significant at the significance level 0.05.
References


Table 1: A small example data set for 9 objects measured on 4 variables (Brusco and Cradit, 2001)

<table>
<thead>
<tr>
<th>Object</th>
<th>Variable 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>15</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>11</td>
<td>7</td>
<td>13</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2: Simulation results for Case 1 and 2 ($p = 3$) given that the number of clusters $K = 3$ is known.

<table>
<thead>
<tr>
<th>$X_3$</th>
<th>$\sigma$</th>
<th>Criterion</th>
<th>Number of variables</th>
<th>Misclassification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(0, \sigma)$</td>
<td>0.1</td>
<td>$G^2$</td>
<td>0.961 0.039</td>
<td>1.80%</td>
</tr>
<tr>
<td>$G^2_0$</td>
<td>1 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>$G^2$</td>
<td>0.851 0.149</td>
<td>1.79%</td>
<td></td>
</tr>
<tr>
<td>$G^2_0$</td>
<td>1 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>$G^2$</td>
<td>0.724 0.276</td>
<td>1.85%</td>
<td></td>
</tr>
<tr>
<td>$G^2_0$</td>
<td>0.990 0.010</td>
<td>1.78%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(X_1 + X_2)/2$</td>
<td>$G^2$</td>
<td>0.382 0.618</td>
<td>2.19%</td>
<td></td>
</tr>
<tr>
<td>$G^2_0$</td>
<td>0.686 0.314</td>
<td>2.05%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Simulation results for the clustering analysis assuming that the number of clusters is unknown for Case 1 and 2 ($p = 3$). Forward stepwise selection procedure combining the number of clusters determination and variable selection is used for the analysis. Variance ratio criterion is used to estimate the optimal cluster size, and the adjusted $G^2_a$ is used for variable selection.

<table>
<thead>
<tr>
<th>$X_3$</th>
<th>$\sigma$</th>
<th>Number of variables</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(0, \sigma)$</td>
<td>0.1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>2</td>
<td>0</td>
<td>0.988</td>
<td>0.001</td>
<td>0</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.009</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$(X_1 + X_2)/2$</td>
<td>2</td>
<td>0.001</td>
<td>0.616</td>
<td>0.062</td>
<td>0.012</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.266</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 4: Simulation results for Case 1 with $p = 100$. When the number of clusters is assumed unknown, forward stepwise selection procedure combining the number of clusters determination and variable selection is used for the analysis where variance ratio criterion is used to estimate the optimal cluster size, and $G^2$ is used for variable selection. This table shows how frequently the two variables $X_1$ and $X_2$ are selected, the number of selected variables ($p_S$), the estimated number of clusters ($K_S$), the misclassification error rate (Mis) and the adjusted rand index (ARI), averaging over 100 simulations. FSS represents the proposed forward stepwise selection method. An item is marked by “-” when it is not needed to be computed.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>Method</th>
<th>Frequency of selecting</th>
<th>$p_S$</th>
<th>$K_S$</th>
<th>Mis</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>All variables, $K = 3$</td>
<td>- -</td>
<td>-</td>
<td>-</td>
<td>0.0189</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K = 3$</td>
<td>1 1</td>
<td>2.2</td>
<td>-</td>
<td>0.0186</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2_a$, $K = 3$</td>
<td>1 1</td>
<td>2</td>
<td>-</td>
<td>0.0181</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K$ is unknown</td>
<td>1 1</td>
<td>2 3</td>
<td>0.0181</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>All variables, $K = 3$</td>
<td>- -</td>
<td>-</td>
<td>-</td>
<td>0.0367</td>
<td>0.897</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K = 3$</td>
<td>1 1</td>
<td>2.75</td>
<td>-</td>
<td>0.0198</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2_a$, $K = 3$</td>
<td>1 1</td>
<td>2</td>
<td>-</td>
<td>0.0183</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K$ is unknown</td>
<td>1 1</td>
<td>2 3</td>
<td>0.0183</td>
<td>0.946</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>All variables, $K = 3$</td>
<td>- -</td>
<td>-</td>
<td>-</td>
<td>0.2640</td>
<td>0.409</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K = 3$</td>
<td>1 1</td>
<td>3.8</td>
<td>-</td>
<td>0.0320</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2_a$, $K = 3$</td>
<td>0.91 0.95</td>
<td>1.86</td>
<td>-</td>
<td>0.0485</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K$ is unknown</td>
<td>0.30 0.56</td>
<td>1.93</td>
<td>3.99</td>
<td>-</td>
<td>0.291</td>
</tr>
</tbody>
</table>
Table 5: Simulation results for Case 2 with $p = 100$. When the number of clusters is assumed unknown, forward stepwise selection procedure combining the number of clusters determination and variable selection is used for the analysis where variance ratio criterion is used to estimate the optimal cluster size, and $G^2$ is used for variable selection. This table shows how frequently the three variables $X_i (i = 1, 2, 3)$ are selected, the number of selected variables ($p_S$), the estimated number of clusters ($K_S$), the misclassification error rate (Mis) and the adjusted rand index (ARI), averaging over 100 simulations. FSS represents the proposed forward stepwise selection method. An item is marked by “-” when it is not needed to be computed.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>Method</th>
<th>Frequency of selecting</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$p_S$</th>
<th>$K_S$</th>
<th>Mis</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>All variables, $K = 3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0235</td>
<td>0.930</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K = 3$</td>
<td>1</td>
<td>0.98</td>
<td>0.56</td>
<td>2.63</td>
<td>-</td>
<td>0.0201</td>
<td>0.940</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2_a$, $K = 3$</td>
<td>0.84</td>
<td>0.89</td>
<td>0.27</td>
<td>2</td>
<td>-</td>
<td>0.0221</td>
<td>0.934</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K$ is unknown</td>
<td>0.76</td>
<td>0.87</td>
<td>0.37</td>
<td>2</td>
<td>3.06</td>
<td>-</td>
<td>0.921</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>All variables, $K = 3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0311</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K = 3$</td>
<td>0.98</td>
<td>0.96</td>
<td>0.56</td>
<td>2.88</td>
<td>-</td>
<td>0.0240</td>
<td>0.928</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2_a$, $K = 3$</td>
<td>0.81</td>
<td>0.74</td>
<td>0.45</td>
<td>2</td>
<td>-</td>
<td>0.0318</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K$ is unknown</td>
<td>0.69</td>
<td>0.68</td>
<td>0.63</td>
<td>2</td>
<td>3.07</td>
<td>-</td>
<td>0.880</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>All variables, $K = 3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0828</td>
<td>0.768</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K = 3$</td>
<td>0.92</td>
<td>0.82</td>
<td>0.40</td>
<td>3.79</td>
<td>-</td>
<td>0.0357</td>
<td>0.895</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2_a$, $K = 3$</td>
<td>0.79</td>
<td>0.72</td>
<td>0.46</td>
<td>1.97</td>
<td>-</td>
<td>0.0448</td>
<td>0.878</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSS, $G^2$, $K$ is unknown</td>
<td>0.46</td>
<td>0.49</td>
<td>0.59</td>
<td>2.14</td>
<td>3.71</td>
<td>-</td>
<td>0.586</td>
<td></td>
</tr>
</tbody>
</table>
Table 6: The $G^2$ and $G^2_a$ values for all the possible subsets of the variables for the cluster analysis of the Brusco & Cradit data (2001)

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$G^2$</th>
<th>$G^2_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5890</td>
<td>0.5890</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.3943</td>
<td>0.3943</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4689</td>
<td>0.4689</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.4689</td>
<td>0.4689</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5890</td>
<td>0.5303</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5187</td>
<td>0.4499</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.5187</td>
<td>0.4499</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.5362</td>
<td>0.4699</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5362</td>
<td>0.4699</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.4689</td>
<td>0.3931</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.5890</td>
<td>0.4520</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5890</td>
<td>0.4520</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.5187</td>
<td>0.3582</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5362</td>
<td>0.3816</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.6076</td>
<td>0.3721</td>
</tr>
</tbody>
</table>
Figure 1: Pairs plot of the Brusco & Cradit data (2001)
Figure 2: A small example data set for 9 objects measured on the two variables $X_1$ and $X_2$ (Brusco and Cradit, 2001): (a) The scatter plot of $X_2$ versus $X_1$; (b) The clustering result using both variables; (c) Project the data into the direction of $X_1$; (d) The clustering result in the direction of $X_1$ only; (e) Assign the memberships of the 2-dimensional data points according to the clustering result based on $X_1$ only; (f) Project the data into the direction of $X_2$; (g) The clustering result in the direction of $X_2$ only; (e) Assign the memberships of the 2-dimensional data points according to the clustering result based on $X_2$ only. Data points in (b), (d), (e), (g), and (h) with the same symbol are assigned to the same cluster, and the filled squares represent the cluster centers.