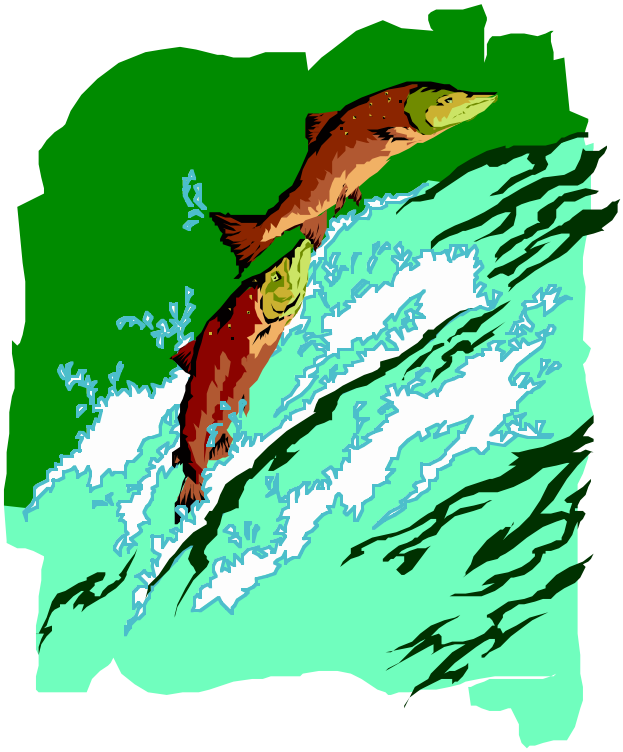


# The Friedman-Rubin approach to Cluster analysis

*(k-means algorithms for the Macintosh)*

Agostino Tarsitano
Dipartimento di Economia e Statistica
Università degli Studi della Calabria
87030 Arcavacata di Rende (Cs)
Italy
Tel. +39-0984-492465
Fax +39-0984-492468

Internet: **agotar@unical.it**

# Contents

# 1. Introduction

The goal of cluster analysis for a given set of data is to verify the presence (or the absence) of natural organization in a fixed number of groups. The data set $D$ consists of $n$ distinct entities $D=\{\boldsymbol{X}_1, \boldsymbol{X}_2,...,\boldsymbol{X}_n\}\subset R^m$ where, for each $r$, $\boldsymbol{X}_r$ gives the observed values of $m$ real-valued characteristics on the entities which are assumed to be known and fixed.

Relative geometric arrangements, causing concentration and dispersion of the entities in different regions, produce clusters.

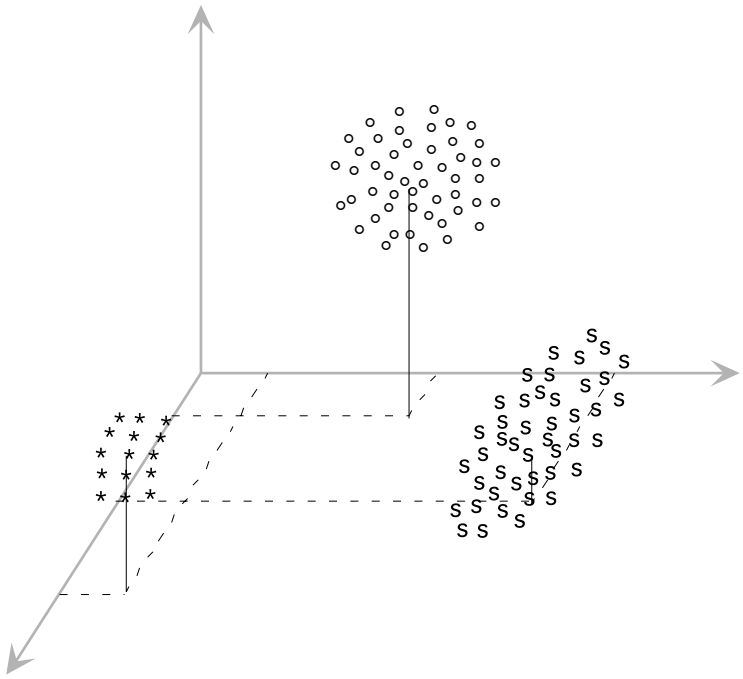


Figure 1: clusters of different shapes and sizes

The figure above depicts a strongly clustered data set consisting of clusters which are homogeneous and well separated. Homogeneity implies that entities in the same cluster are near each other. Separateness implies that entities in different clusters are farther one from the other.

The entities are unlabeled. All we have is a collection of vectors associated with a given set of variables without knowing if the entities belong to different categories, if there is more than one category and the category membership of the entities included in the data set.

## 1.1 Overview

To learn something from such an unpromising basis depends upon the assumptions one is willing to accept. Suppose that the entities came from a distribution for which the multivariate second moments exist. Then a compact description of the data set can be obtained by the sample mean and the sample covariance matrix.

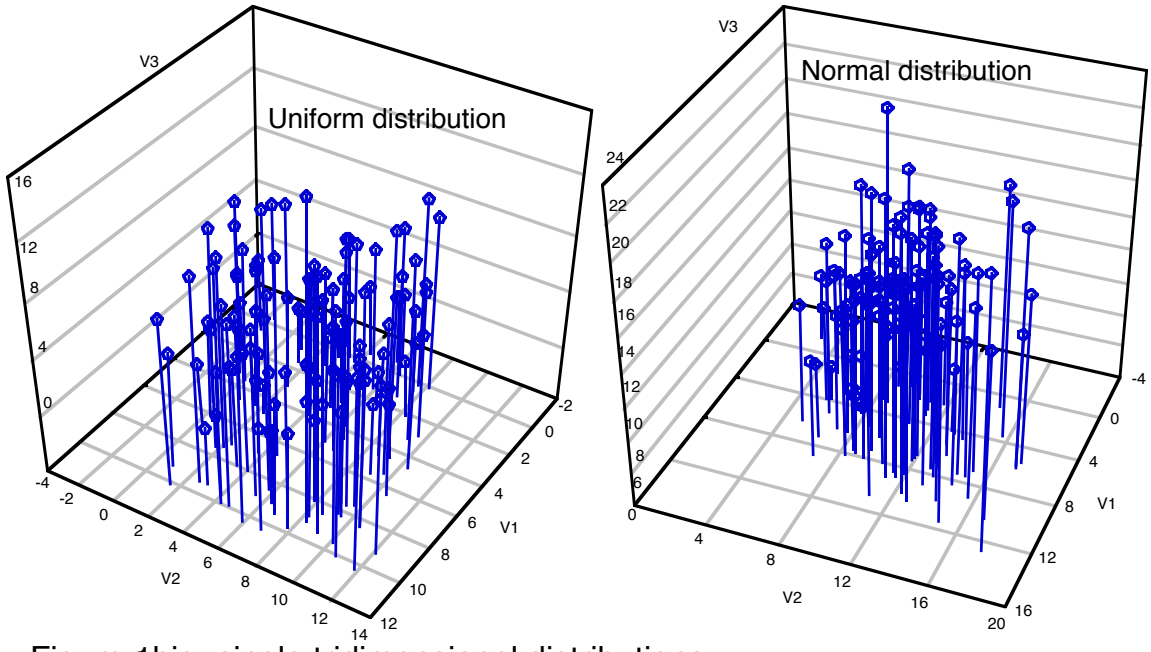



Figure 1bis: single tridimensional distributions

The two graph above represent a sample *of n=120* entities from, respectively, a uniform and a normal 3-dimensional distribution having $\mu=(5,10,15)$ *and* $\sum=(\sigma_{ij})$, $\sigma_{ii}=10$, $\sigma_{ij}=5$ *for* $i\neq j$. In general, second-order statistics are incapable of revealing all of the structure in a set of data since other distribution may differ for other important features then mean and covariance. If we assume that entities fall in hyperellipsoidally shaped clouds we can approximate a great variety of situations.

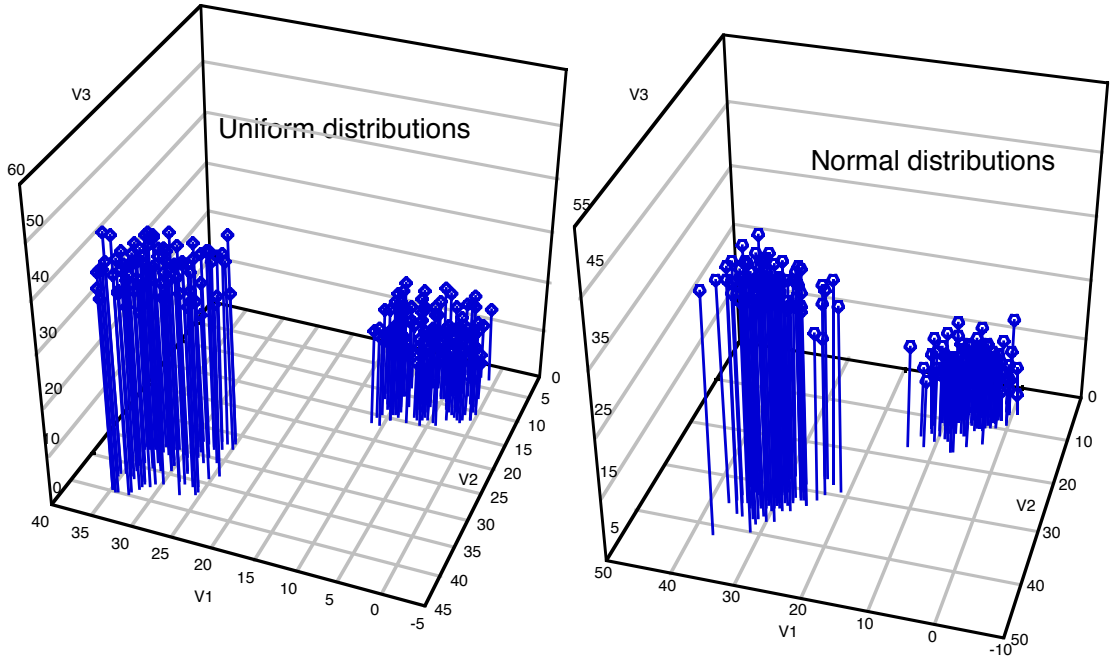



Figure 1ter : two-group tridimensional distributions

Fortunately, the type of approximation we are looking for is not hard to please. The only thing that must be learned is the values of an unknown parameter vector which maps the set of entities to the set of group labels. Figure 1ter illustrates the problem. The two clusters in both graphs have different means and different variance-covariance matrices. If the normal distribution is used to approximate the uniform distribution the results can be very misleading. But, this is not the case. The normal distribution is used for an easier task: distinguishing the entities which fall into the first cluster from the entities falling into the second cluster and this can be fully accomplished even if the approximation of the vector means and the variance-covariance matrices are poorly estimated.

The present paper assumes that the data set has clusters which tend to take the form of hyperellipsoid of various size, but with the same orientations which is the essence of the Friedman-Rubin approach to cluster analysis.

A brief outline of my method of working will help explain the contents of the article. The rest of this section reviews the problem of assessing the partitional adequacy of the subdivision into a fixed number of groups of a data set.

In section 3 a new method for relocative scheme which minimizes the determinant of the sample within-group dispersion matrix is proposed and tested by looking at various real and simulated applications. The main difference with other k-means is a transfer technique which realizes a global best step instead of a local best step.

Section 3 describes the initialization methods. Section 4 outlines the stopping rules and studies their shortcomings and merits in connection with problems arising in practical applications. The software which implements the algorithms is described in section 5.

### 1.2 The partition of a data set

Seber (1984, p. 379) stated that the major weakness of agglomerative hierarchical methods is the constrain that the (k+1)-partition must be included in the k-partition so that an improper fusion at an early stage cannot be corrected later. On the contrary, the essence of a k-means algorithm is the relocation of entities which gives these techniques an immense advantage over the others. This section reviews the general framework of the k-means algorithm for the subdivision of the data set into a fixed number of mutually exclusive and exhaustive clusters. A partition (or a clustering) $\gamma$ of a finite set of n entities $D=(\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_n)$ is a collection of $k$ subsets, called the clusters of $\gamma$, such that

$$C_j \neq \varnothing,\ 1 \leq k \leq n;\quad \bigcup_{j=1}^{k} C_j = D;\qquad C_i \cap C_j = \varnothing \quad (\ i \neq j\ );\tag{1}$$

where $\varnothing$ is an empty set. The cardinalities $n_1, n_2, \ldots, n_k$ of the clusters satisfy

$$a)\ 1 \le n_i \le n-k+1,\ i=1,2,\ldots,k \qquad b)\ \sum_{i=1}^{k} n_i = n \tag{2}$$

This implies that each entity is assigned to one cluster, each cluster contains at least one entity and the partition contains all entities.

A partition can be succinctly expressed by the classification vector $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_n)$ which maps the set of entities to the set of cluster labels $\boldsymbol{X}_r : \gamma_r = j$ if $\boldsymbol{X}_r \in C_j$.
The number of clusters $k$ is assumed to be given as input, although it is often unknown and its estimation is a topical problem in cluster analysis. The program, as such, is not able to merge small clusters that are very close and no larger clusters are broken up. The approach suggested by ***DetClus*** is to run the program for a range of values of $k$

$$2 \le k_1 \le k_2 \le (n-k) \tag{3}$$

and to empirically determine the best number of clusters. The upper (lower) limit for $k$ should be at least 3 or 4 more (less) clusters than are ultimately suspected (these limits are necessary since, if $k$ is excessively large or small, spurious or unnatural clusters tend to appear). For each $k$ the program carries out the clustering regardless of the previous grouping and computes a series of clustering quality indices which allow the user to decide the appropriate value (or values) for the number of clusters.

### 1.3 The definition of a cluster

It is almost a commonplace that there exists no agreed upon idea of a cluster and that, according to the scope of the analysis, different type of clusters are allowed (a typical example of such vagueness is the distinction drawn between natural and arbitrary clusters proposed by Kruskal (1977):" ... *We call clusters natural if the membership is determined fairly well in a natural way by the data, and we call the clusters arbitrary if there is a substantial arbitrary element in the assignment process*".

The k-means approach to cluster analysis is based on a "metric" concept of a cluster. The $n$ entities are confined to an m-dimensional parallelepiped

$$R = \left\{ X | X_i \in \left[ x_i^{'}, x_i^{''} \right];\ i = 1,2,\ldots,m \right\}; \qquad with \quad x_i^{'} = \underset{1 \le r \le n}{Min} \left\{ x_{ri} \right\},\ \ x_i^{''} = \underset{1 \le r \le n}{Max} \left\{ x_{ri} \right\} \tag{4}$$

Clusters are accumulations of points in distinct regions of *R* entirely surmounted by empty space (see figure 2).

Cluster analysis techniques search partitions characterized by remoteness in which, as was observed by Cormack (1971), two conflicting requirement are involved: internal compactness or cohesion (i.e. object belonging to the same cluster are in some operational sense similar to each other) and external isolation (very dissimilar entities must be placed in different clusters). The two factor are dependent: a highly dense accumulation of points (A) needs less isolation to be considered a proper cluster, and, sometimes, a very sparse group (B) is accepted as a single cluster only for the substantial gap between its entities and the others. The size of the clusters is also important: internal homogeneity tends to be greater for small clusters than for large ones; external isolation has an opposite tendency. According to Ling (1973) a cluster is judged real if it is significantly compact or isolated or both.

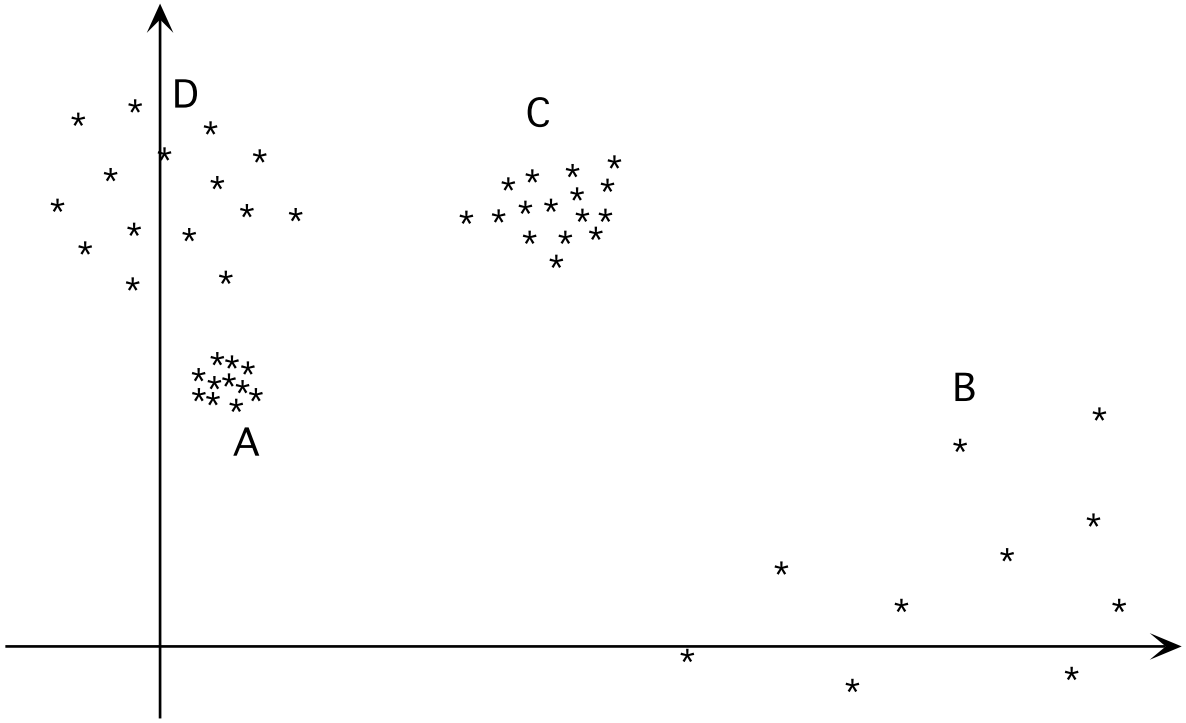



Figure 2: compactness and isolation of clusters

The concept of remoteness is vague as almost is everything referred to in cluster analysis. For instance it is not clear how to deal with the disturbing (but inescapable in real applications) phenomenon of intermediate entities (borderline or hybrid clusters) linking two cohesive and otherwise isolated clusters.

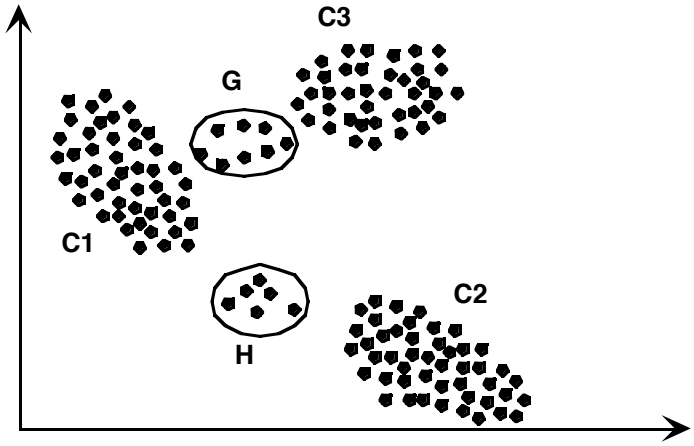



Figure 3a: hybrid clusters

The cluster G in Figure 3a is an object which depicts the transformation process followed by the entities in passing form status C1 to C3 (or *vice versa*). The cluster H is a structure formed by the entities which share the characteristics of both C1 and C2. If the clustering algorithm is such that the entities in the chain cluster G and hybrid cluster H have to be assigned to one of the two major cluster, C1 and C2, their isolation is doubtful.

Cluster Y rises another question. Is such structure shared by enough entities to be considered real and worthy of attention or is it a mere product of random turbulences in data collection? No easy answer exists.

Another somewhat undesirable phenomenon is the presence of outliers or singletons (Figure 3b) that is, clusters formed by a single entity whose distances from the other (n-1) entities are all significant. What is the correct number of clusters? Three (ignoring X), or four (considering X a genuine cluster)? If one considers X a unique case which does not reserve further treatment then the clustering algorithm can run on a reduced dataset from which X has been eliminated. This has the advantage of limiting the number of contenders to whom an entity could be assigned. If X cannot be discarded then this has dramatic effects on the general lookups of the clustering.

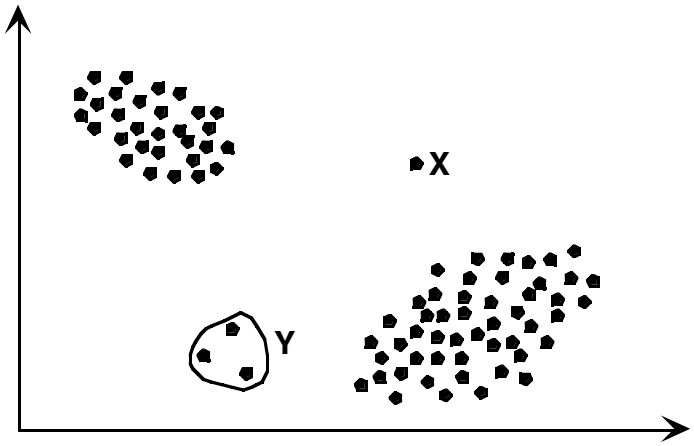



Figure 3b: a singleton and a small cluster

In this sense, the requirement of exclusive assignment 2b is particularly strong because every entity is forced to join a cluster whereas one would ordinarily be inclined to separate out outliers (entities which fits badly into existing clusters) or being just naive or intermediate entities linking two or more otherwise isolated clusters. Applications of k-means to real data should be able to handle "nuisance" entities from further clustering runs (Bayne *et al*, 1980), although such question represents an important research challenge. Since some clustering criterion is very sensitive to the presence of outliers, some attempt should be made to remove these. It is clear that choices made at this stage can have a determining influence on the output of the subsequent analysis.

### 1.4 Perfect and admissible clustering

Clustering methods have a common intuitive requirement: entities in the same cluster should be closer than entities in different clusters. Rubin (1967) called "well-structured" such partitions. Ideally (Rubin, 1967; Van Rijsbergern, 1970) one would ask that the maximal distance between entities in the same cluster be lower than the minimal distance between entities in different clusters.

Let

$$\Delta_j = \underset{\gamma_r=j,\gamma_s=j}{Max} \left\{ d(\boldsymbol{X}_r, \boldsymbol{X}_s);\ r,s = 1,\ldots,n \right\};\ S(i,j) = \underset{\substack{\gamma_r=i,\gamma_s=j \\ i\neq j}}{Min} \left\{ d(\boldsymbol{X}_r, \boldsymbol{X}_s);\ r,s = 1,\ldots,n \right\} \tag{5}$$

denote, respectively, the diameter and the moat of the cluster $C_j$. Two type of ideal clustering can be defined. The first is the “perfect clustering”

$$\Delta_j \ \leq \ S_j \quad j = 1,2,\ldots,k \tag{6}$$

An example of perfect clustering is the disjoint partition consisting of each entity in a separate cluster. In this case $\Delta_j = 0,\ j=1,2,\ldots,k$ and $S_j > 0,\ j=1,2,\ldots,n,$ supposing that the X's are distinct. Nonetheless, perfect clustering is too restrictive (Bailey and Dubes, 1982; Tarsitano and Anania, 1995) since it eliminates many reasonable groupings.

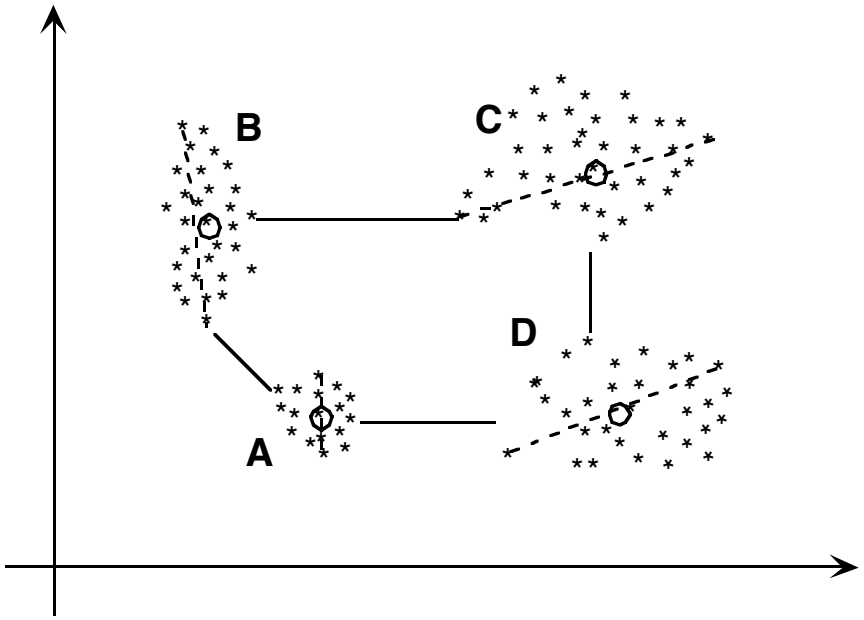



Figure 4: Admissible, but non perfect clustering

A less stringent definition, also useful from a computational point of view, is the string condition (Rao, 1971): in an admissible clustering, each group consists of at least one entity $\mu_j$ such that the distance between it and any entity that does not belong to the same cluster is not less than the distance between $\mu_j$ and any entity within the same cluster.

$$\underset{\gamma_r \neq j}{Min}\left\{d\left(X_r, \mu_j\right);\ r = 1,\ldots,n\right\} \geq \underset{\gamma_r = j}{Max}\left\{d\left(X_r, \mu_j\right);\ r = 1,\ldots,n\right\};\qquad j = 1,2,\ldots k \tag{7}$$

According to this definition (see Figure 4), the problem addressed by k-means algorithms is to discover, for each cluster, a representative or typical entity for whom is minimized a known function of the dissimilarities between an entity in the cluster and the centroid.

The centroid can be either a hypothetical entity which is not an entity in the cluster (*e.g.* the vector of the arithmetic mean of all entities currently in the cluster) or an existing entity (*e.g.* the most typical entity that is the entity with the smallest average or total distance between itself and the other entities in the cluster). When centroids are defined the classification vector $\gamma$ is determined by assigning all entities to the most similar centroid.

The ability of a centroid to summarize the information content of the cluster depends on the actual spread of the data in the given variable space. Usually, the centroids, the cluster membership, and the variance-covariance structure are unknown and must be estimated from the data set. Since each partition provides an estimate of the parameters, some selection is necessary. A comparison can be accomplished using an objective function $L(\gamma):\gamma \in P(n,k) \rightarrow [0,\infty)$ (here $\gamma_r$ denotes the cluster membership assigned to $\boldsymbol{X}_r$) such that $L(\gamma)<L(\delta)$ means that $\gamma$ provides better estimates than $\delta$ (where $P(n,k)$ denotes the set of partitions of $n$ entities into $k$ clusters). Since the cardinality of $P(n,k)$ is finite, it exists at least one partition $\gamma^*$ such that

$$L\left(\gamma^*\right) = \underset{\gamma \in P(n,k)}{Min}\left\{L(\gamma)\right\} \tag{8}$$

The most straightforward way to find $\gamma^*$ is to evaluate $L(\gamma)$ for all $\gamma \in P(n,k)$. It is well known, though, that this is not a viable solution since the cardinality of $P(n,k)$ grows rapidly (it is of the order $k^n/k!$) and becomes prohibitively high even for moderate values of $n$ and $k$. The k-means partitioning is a "NP-hard problem" for which no a priori guarantee can be given in terms of solution quality and running time. Although the dividing line between things which are practical to compute and things which are not is continuously pushed forward, the search for $\gamma^*$ must still be conducted over a small subset of $P(n,k)$ using strategies which find solutions that are often good, but not necessarily optimal.

## 2. Characteristics of a k-means algorithm

There is a wide choice of clustering methods which have different adaptability to the data and different requirements of computer resources. Given an initial partition $\gamma^q$ with $q=0$, k-means algorithms compute the criterion value $L(\gamma^q)$. Another partition $\gamma^{q+1}$ is obtained by transferring a single entity (or block of them) from one cluster to another. The transition from $\gamma^{(q)}$ to $\gamma^{(q+1)}$ can be realized by means of up- and down-dating formulae for the exchange of entities between clusters. The new partition is accepted if $L(\gamma^{q+1})<L(\gamma^q)$ and the procedure is repeated until no further reduction of $L(\gamma^q)$ can be obtained.

The algorithm terminates after a finite (typically small) number of iterations. It is worth pointing out that the k-means partitioning is a "NP-hard problem", that is, there is no absolute guarantee in terms of solution quality and running time.

There are many variations of, and extensions to, this approach and the lack of investigations into their properties is, in large measure, due to the excess of options which form an k-means algorithm. The ***DetClus***, as with any program implementing a relocation scheme, has the following essential phases:

0) Feature selection
1) Determining the criterion (distance measure)
2) Starting the process
3) Reassigning entities
    3.1 Distribution of the entities among clusters
    3.2 Updating the centroids, cardinalities and scatter matrices
4) Overcoming local minima
5) Validation of the results
6) Interpretation of the results

The variables must be properly chosen so as to englobe as much information as possible concerning the difference between entities, but, with the minimum number of uncorrelated features. These issues are, however, outside the scope of the present section.

### 2.1 Determining the criterion

Iterative schemes are concerned with making membership changes which optimize a numerical criterion. The choice of the objective function $L(\gamma)$ is crucial for a K-means algorithm because it must take into account two requirement which may be difficult to reconcile. One goal (internal cluster cohesion) can conflict with another (separation between clusters).

Several criteria have been proposed and each of them is predisposed to finding certain type of clusters and has specific properties. ***DetClus*** is based on the criterion proposed by Fried-

man and Rubin (1967)

$$L(\gamma) = Min\left\{\left|\boldsymbol{W}_{\boldsymbol{q}}(\gamma)\right|\right\}$$

$$\boldsymbol{W}_q = \sum_{j=1}^{k} \boldsymbol{W}_j^q; \quad \boldsymbol{W}_j^q = \sum_{r=1}^{n} \left(\boldsymbol{X}_r - \mu_{\gamma_r}^q\right)\left(\boldsymbol{X}_r - \mu_{\gamma_r}^q\right)^t; \; \mu_j^q = \sum_{\gamma_r = j} \boldsymbol{X}_r \Big/ n_j^q \;, \; j = 1,2,\ldots,k \tag{9}$$

Where $\boldsymbol{W}_q$ is the pooled dispersion matrix across the $k$ clusters (or "within-group" dispersion matrix) for the q-th classification vector. In order for (9) to be non-singular, it is required that $(n-k) \geq m$ otherwise the estimate is singular regardless the true value of **W**. Naturally, since total dispersion matrix $\boldsymbol{T}$ is fixed for every partition of the given data set, $Min\{|\boldsymbol{W}(\gamma)|\}$ is equivalent to $Max\{|\boldsymbol{T}|/|\boldsymbol{W}(\gamma)|\}$. A simple variation of (9) was proposed by Symons (1981)

$$L\left[\gamma_q\right] = nLn\left[\left|\boldsymbol{W}\left(\gamma_q\right)\right|\right] - 2\sum_{j=1}^{g} n_j^{(q)} Ln\left(n_j^{(q)}\right)$$

Some empirical results does not support such criterion since relocations based on it stop after surprisingly few iterations.

The minimization of the determinant of $\boldsymbol{W}(\gamma)|$ does not make such restrictive assumptions about the shape of the clusters as does $Min\{\mathrm{Tr}[\boldsymbol{W}(\gamma)]\}$ assuming only that the clusters has the same shape and orientation, but not that they are spherical. Although computationally more involved and expensive, the criterion $Min\{|\boldsymbol{W}(\gamma)|\}$ is invariant under the affine transformations $\boldsymbol{Y}=\boldsymbol{AX}+\boldsymbol{b}$ where **A** is non singular (this allows the question of standardization of the variables to be overcome and the results do not depend on arbitrary factors such as the units of measurement used for data acquisition). Furthermore, it reduces the repetitive effect of several highly correlated attributes by considering sums of cross products in addition to sums of squares (Arnold 1979).

The use of (9) implies that the dissimilarities between the entities are measured by the generalized (Mahalanobis) distances, each centroid coincides with the averages of all entities within the cluster and the clusters have the same variance-covariance matrix. In fact, Mahalanobis distances are equivalent to the Euclidean distances between the transformed entities:

$$\boldsymbol{Y}_i = \boldsymbol{H}\boldsymbol{X}_i \;, \qquad i=1,2,\ldots,m$$

where $\boldsymbol{H}\boldsymbol{H}^t$ is the Cholesky factorization of $\boldsymbol{W}$.

The generalized distance introduced by Mahalobis (1936) is a distance measure corrected in terms of the group structure of the data. Additionally, it is appropriate when the variables are

correlated because it takes into account the variability of the values in all dimensions. The point C in figure 5, which clearly lies in the domain of cluster B would be allocated in cluster A if the Euclidean metric were used. If the within-cluster covariance matrix is known, the data can be transformed $\boldsymbol{Y}_i = \boldsymbol{H}\boldsymbol{X}_i$ to make the clusters spherical so that the Euclidean distance can be used. But when we are doing a cluster analysis, we do not know what the true cluster membership is and we cannot calculate W so that an approximation should be used.

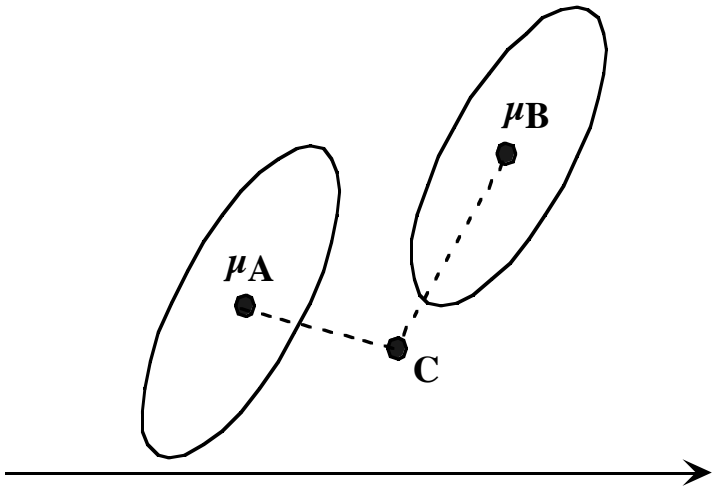



Figure 5: Euclidean vs Mahalanobis distance

Since

$$|\boldsymbol{W}(\gamma)| = \prod_{i=1}^{m} \lambda_i(\gamma)$$

where $\lambda_i$ is the i-th eigenvalue of the within-cluster scatter matrix, the criterion $Min\{|\boldsymbol{W}(\gamma)|\}$ tries to minimize the volume of the hypercube defined by the variances in the direction of the $m$ principal axes of the data set. This means that (9) is appropriate when the variables are correlated because it takes into account the variability of the values in all dimensions. However, since the within-group dispersion matrix $\boldsymbol{W}$ is an average of the variance-covariance matrices of the clusters, correlated variables in the clusters generate multicollinearity in $\boldsymbol{W}$. In other words, $|\boldsymbol{W}|$ will approach to zero as correlations grow stronger. The Mahalanobis distance between the centroids, calculated by using $\boldsymbol{W}^{-1}$, tends to infinite; as a consequence, the only clustering compatible with such conditions is the disjoint partition. However, since the within-group dispersion matrix $\boldsymbol{W}$ is an average of the variance-covariance matrices of the clusters, correlated variables in the clusters generate multicollinearity in $\boldsymbol{W}$. In other words, $|\boldsymbol{W}|$ will approach to zero as correlations grow stronger. The Mahalanobis distance between the centroids, calculated by us-

ing $\boldsymbol{W}^{-1}$, tends to infinite; as a consequence, the only clustering compatible with such conditions is the disjoint partition.

In the general case, when the centroids $\mu_i$, $i=1,2,...,k$ and the matrix $\boldsymbol{W}$ are completely unknown, the number of unknown parameters to be estimated equals $km+m(m-1)/2$ and, therefore, reliable estimates are possible only if $n$ is much more greater than this threshold. If $\boldsymbol{W}$ is ill-conditioned and one supposes that the $k$ clusters lie in the same subspace, redundant features can be eliminated by representing the data in a new coordinate system in which the effective description can be given by applying techniques to reduce the dimensionality of the data. An evident technique is to apply Principal Component Analysis and to perform the cluster analysis on the factor scores of the first few leading factors instead of the complete data (the use of PCA as well as factor analysis is contraindicated if each variable is endowed of a useful and independent discriminating power). While this can be helpful for finding clusters it can make results difficult to interpret.

Dimension reduction has, however, many positive implications. Firstly, for the computational effort because reduced data require less storage space and can be manipulate more quickly than the original set of variables. Secondly, a limited set of selected features may alleviate the influence of irrelevant information (features showing little differentiation across the data set or highly correlated with other features) to whom the Mahalanobis norm give the same relative importance as the other variables thus degrading the grouping ability of the most salient features. Third, to avoid implicit weighting: if two collinear features are used, then their common dimension is effectively double weighted (Heeler and Day, 1975). In addition, eliminating redundant variables helps to interpret and compare the configurations derived by cluster analysis. Finally, some validation tests (*e.g.* the $C^3$ clustering criterion) designed for uncorrelated variables, becomes applicable to orthogonal principal components.

*Example 1:*

Economics of Cities. (http://lib.stat.cmu. edu/ DASL/).

The data represent the economic conditions in 46 cities around in world in 1991. The variables are: work (weighted average of the number of working hours in 12 occupations), price (index of the cost 112 goods and services excluding rent, Zurich =100), salary: (index of hourly earnings in 12 occupations after deductions (Zurich =100). If all the PC's are used for the calculations the Mahalanobis distances between the point of the figure 6b are equivalent to the Euclidean distance between the points of Figure 6a.

However the appearance of the data sets in the normalized PC space is different from the original space, since now, along each PC, the points have the same variance.

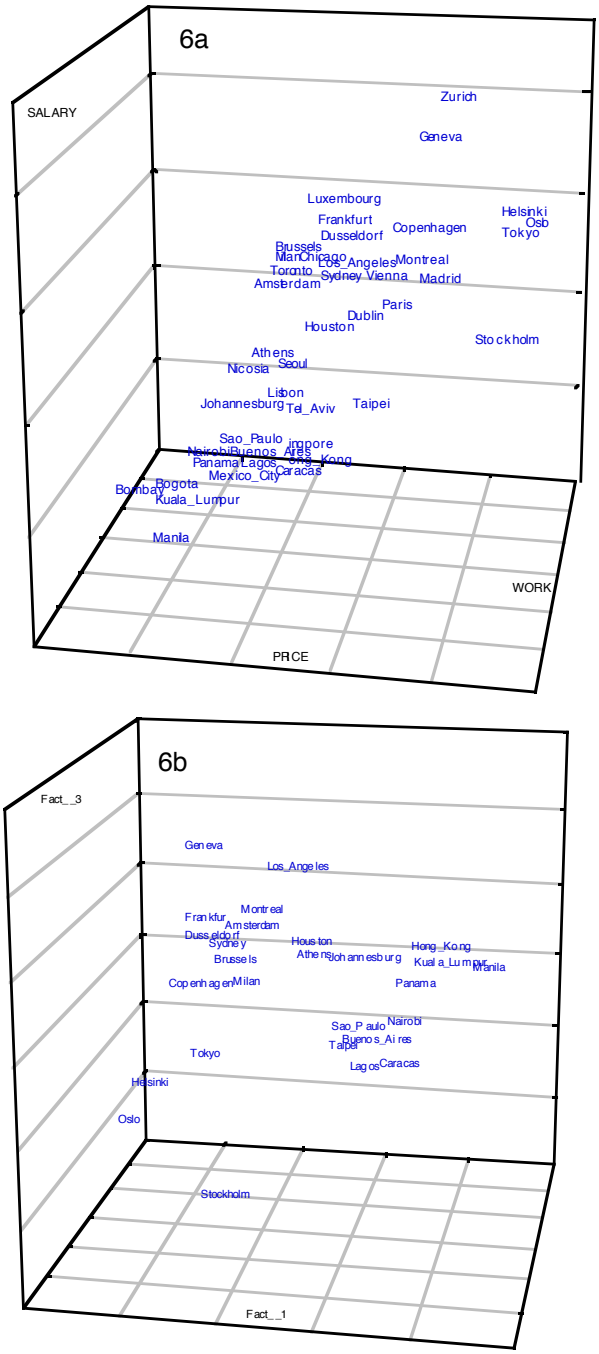



Figure 6: cities data set.
Original an normalized PC spaces

***Example 2:***

Sexual activity and the lifespan of male fruitflies. Source: "Sexual Activity and the Lifespan of Male Fruitflies" by Linda Partridge and Marion Farquhar. Nature, 294, 580-581, 1981. Size:125 observations, 5 variables: number of partners (0, 1 or 8), Type of companion: 0=newly pregnant female; 1= virgin female:9: not applicable (when partners=0), lifespan, in days, length of thorax, percentage of each day spent sleeping. The first two variables are used as if they were quantitative, although is questionable as to how far these variables can be c as metric.

The pooled within-cluster scatter matrix is singular for any value of $k$ and criterion (9) cannot be applied to this data set. However, the the first four principal components of the correlation matrix (explaining the 93.2% of the total variation) indicate the presence of a group structure although the number of clusters is uncertain. ***DetClus*** run plainly on the reduced data set ensuring a perfect revovery of the five cluster present in the data.

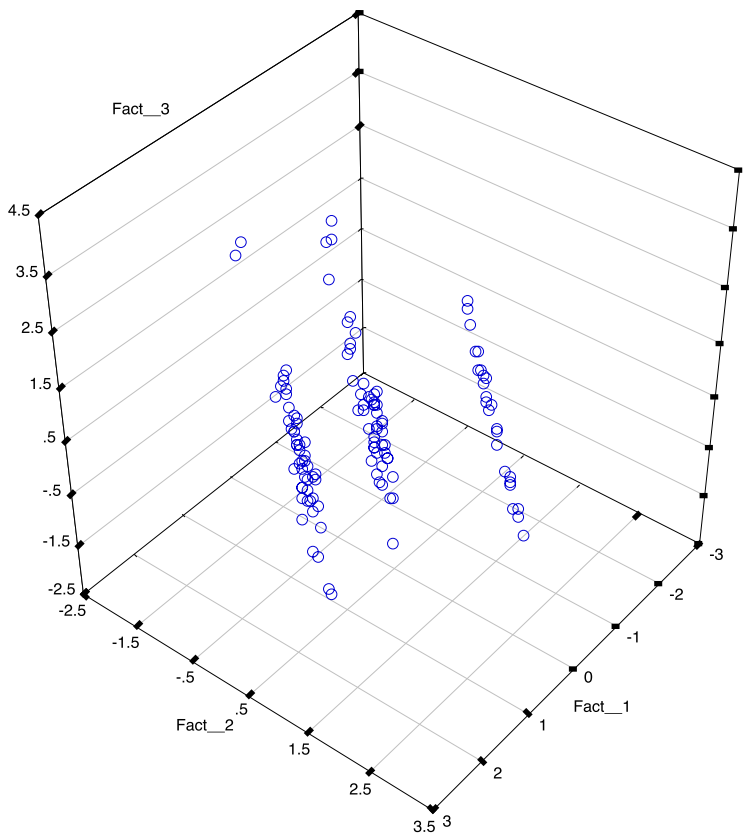



Figure 7: fruitfly reduced data set

## 2.2 Interpretation of the criterion

Friedman and Rubin relate *Min{|W|}* to Wilks' lambda statistic encountered in the multivariate analysis of variance. In this context, the hypothesis that the means of $k$ normal multivariate distributions with a common dispersion matrix are equal $H_0$: $\mu_1=\mu_2= ...=\mu_k$ versus $H_1$: *at least one* $\mu_i \neq \mu_j$ *is available* is tested by considering

$$\boldsymbol{T} = \sum_{r=1}^{n} (\mathbf{X}_r - \mu)(\mathbf{X} - \mu)^t = \sum_{r=1}^{n} \left(\boldsymbol{X}_r - \mu_{\gamma_r}\right)\left(\boldsymbol{X}_r - \mu_{\gamma_r}\right)^t + \sum_{j=1}^{k} n_j \left(\mu_j - \mu\right)\left(\mu_j - \mu\right)^t = \boldsymbol{W} + \boldsymbol{B} \tag{10}$$

where $\mu$ is the total mean and $\boldsymbol{B}$ is the "between" dispersion matrix. Specifically, $H_0$ is rejected if $\lambda=|\boldsymbol{W}|/|\boldsymbol{W}+\boldsymbol{B}|$ is too small. Since $\boldsymbol{W}+\boldsymbol{B}$ is fixed, minimizing the determinant of the within dispersion matrix is equivalent to minimizing the p-value of the Wilk's lambda.

The minimization of $|\boldsymbol{W}|$ searches for clusters that are hyper-ellipsoidal with equal orientations. Everitt (2001), Chernoff (1970), Chen *et al*. (1974), Symons (1981) points out that the metric $\boldsymbol{W}$ could produce incorrect and misleading results when the dispersion structures of the clusters are markedly heterogeneous. In fact, the algorithm is destined to find football-shaped clusters sharing a common orientation even if there were no trace of them in the data set.

*Example 3:*

Diday and Govaert's data. Fifty observations from each of three bivariate normal distributions, as described in Diday and Govaert, RAIRO Informatique/Computer Sciences, 11, 329-49 (1977). The data have been studied also in Gordon (1999, 46-48). The phenomenon is clearly illustrated in Figure 8 where the minimization of the determinant has driven the algorithm along the axis of maximal dispersion.

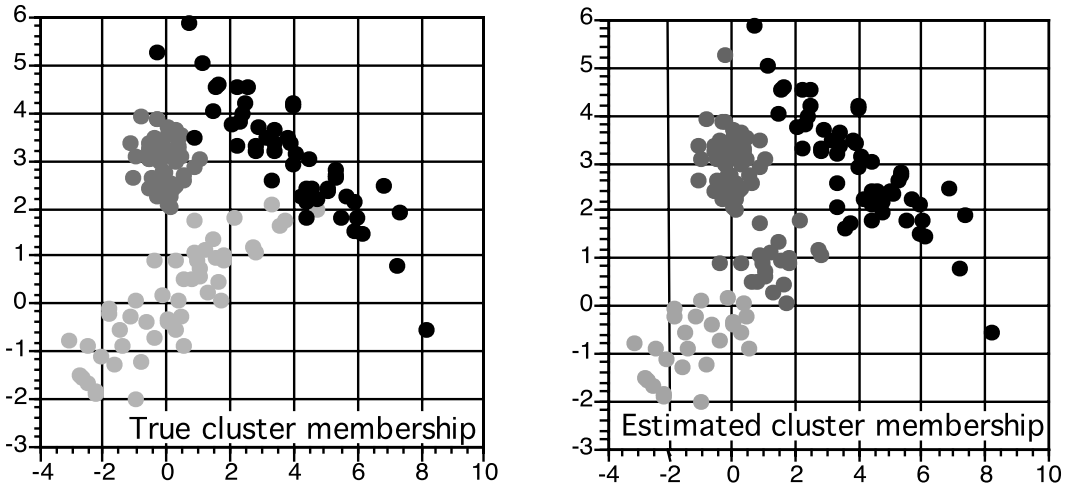



Figure 8: peculiarity of Min{|W|}

Several authors have pointed out that if the condition of homogeneity for the within-group dispersion matrices is not satisfied, the clustering based on $Min\{|\boldsymbol{W}|\}$ may impose artificial structure which precludes uncovering patterns hidden below the surface of data

Scott and Symons derived the criterion within the context of the maximum likelihood estimation of γ assuming a multivariate normal distribution for each clusters. Bayne et al. (1980) found that $|\boldsymbol{W}|$ and $Tr(\boldsymbol{W})$ do not differ significantly. Zemroch (1996) notes that the Mahalanobis distance is attractive because it emphasizes the unusualness of those entities that most defy the intrinsic relationships among the variables. Marriott argues that if one of the variables is strongly grouped, the minimum partition defined by (3) would be entirely based on that variable. Such a partition obeys the string condition, but the clusters might be unrecognizable if viewed against the entire space of the variables. Scott and Symons (1971) conjecture that the criterion encourages the formation of partitions with clusters of equal size which is not necessarily a curse, especially in many experimental designs. Exhaustive experimentation with $|\boldsymbol{W}|$ used jointly with an efficient procedures (see section 3) for determining the initial configuration, does not confirm such a tendency, at least for well structured data set.

*Example 4:*

Fishcatch data set (available from the data archives of the Journal of Statistical Education). A sample of 157 fishes of 7 species are caught and measured. All the fishes are caught from the same lake (Laengelmavesi) near Tampere in Finland. The best solution for k=7 and k=5 are shown in figure 9. In the first case the data are weakly clustered and the tendency to nearly equal sized clusters 8 column "E") is confirmed. For $k=5$ the group structure is more pronounced and the clustering based on $Min\{|\boldsymbol{W}(\gamma)|\}$recognizes both small and large clusters.

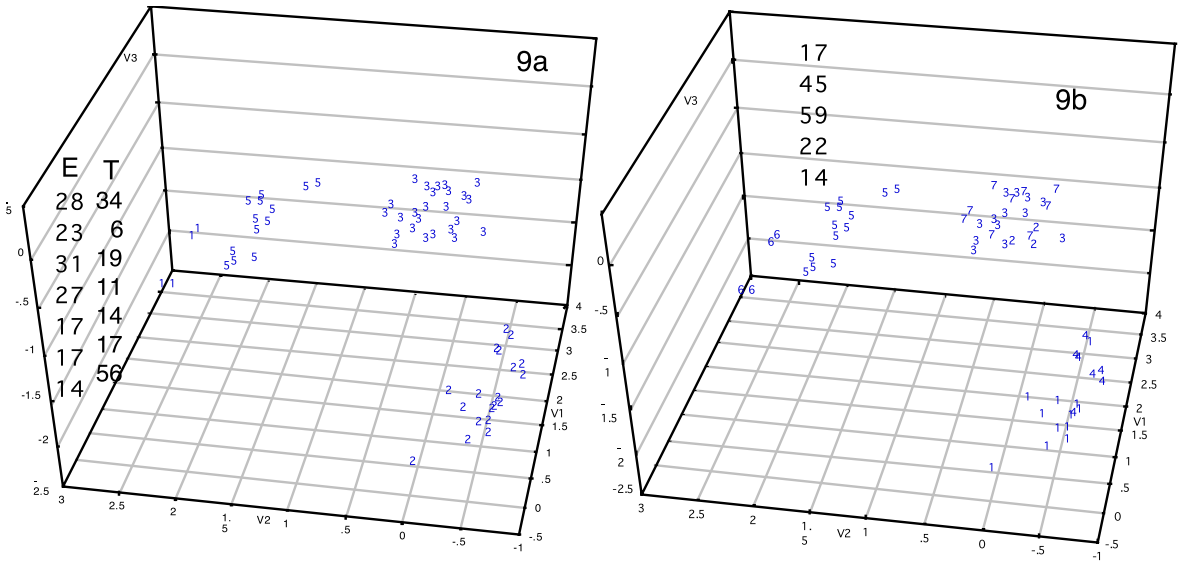
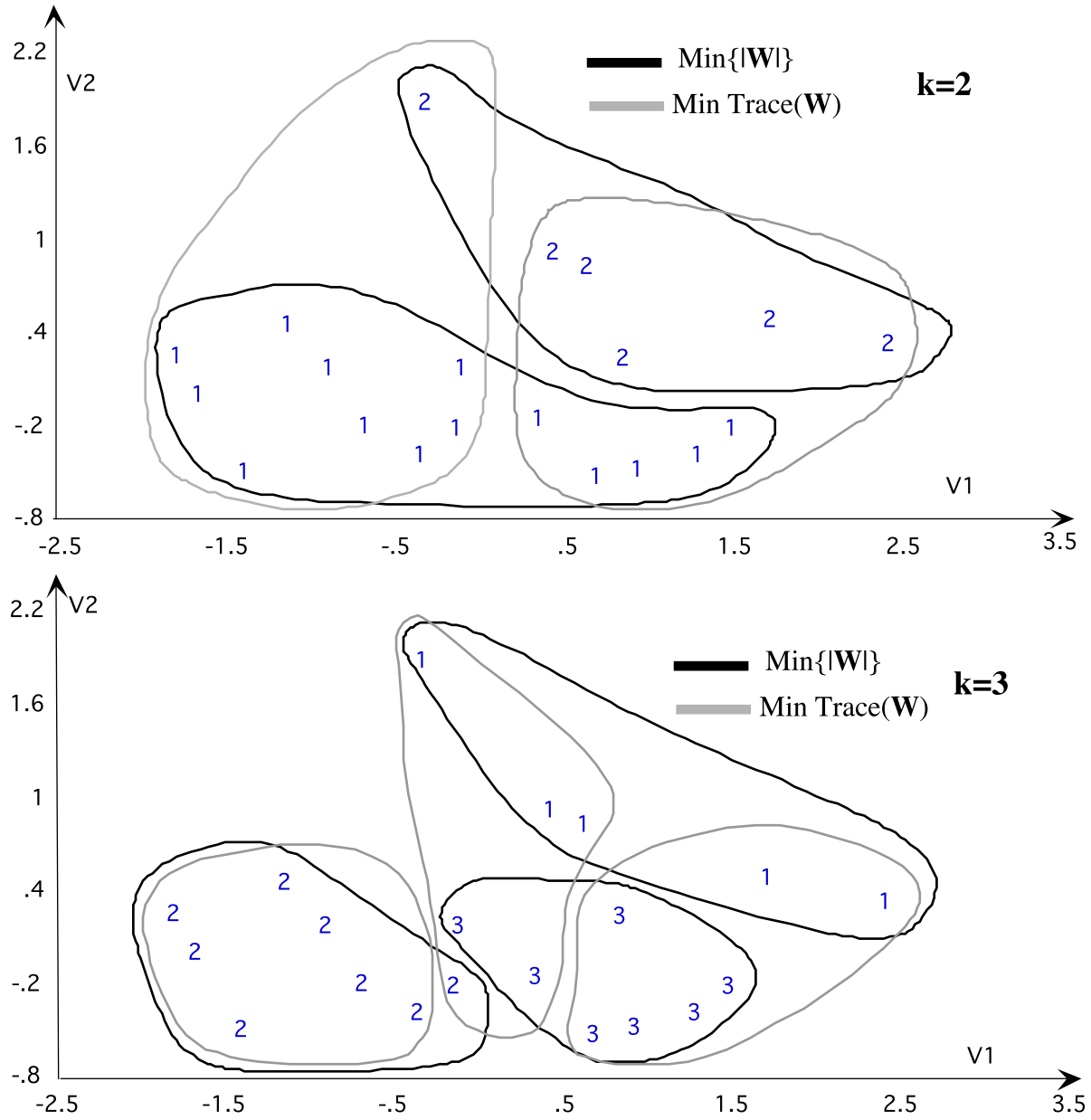



Figure 9: best solutions for k=7 and k=5

***Example 5:*** Duda *et al*. (2001, pp. 543-548) discuss various criterion functions fro clustering by applying the criteria to a simple data set. The raw data does not exhibit any obvious clusters.




For *k=2* the clusters found by minimizing the sum os squared errors (*Tr(**W**)*) tend to favor clusters of roughly equal number of entities; in contrast, *Min|**W**|* favors one large and one fairly small cluster (Bayne *et al.* 1980 found that |***W***| and *Tr(**W**)* do not differ significantly).The clusters in the figures are stretched horizontally because the variation of the data set is greater along the V1 axis than along the V2 axis (the solution found by **Detclus** is different from that presented by Duda *et al*. 2001). For *k=3* the difference between the clusters determined by the two criteria becomes smaller (the first cluster on the left is almost the same). According to *Duda et al*. this is a general tendency.

### 2.3 Reassigning entities

The essence of a k-means algorithm is the reallocation phase and, in fact, the type of pass is a distinctive feature of the method. There are a number of schemes in common use to relocate entities, each reflecting a different trade-off between classification capability that can be achieved and computer time consumed. Most methods differ basically in the number of criterion evaluations required to reach a minimum and the accuracy of this minimum.

The schemes considered by ***DetClus*** are based on a combination of two distinct stages: transfers and swaps. Transfers consist of moving one entity from one cluster to another; swaps involve the exchange of two entities from different clusters.

Let $|\boldsymbol{W}_{q+1}|$ the determinant of the within-cluster dispersion matrix after that the transfer of $\boldsymbol{X}_r$ from cluster $j$ to $i$ has taken place (the transfer from a singleton is not considered).

$$\begin{aligned}
&\Delta_q(r,j,i) = \frac{\left|\mathbf{W}_{q+1}\right|}{\left|\mathbf{W}_q\right|} = \left(1+\alpha_i \mathbf{y}_i^t \mathbf{W}_q^{-1} \mathbf{y}_i\right)\left(1-\alpha_j \mathbf{y}_j^t \mathbf{W}_q^{-1} \mathbf{y}_j\right) + \alpha_i \alpha_j \left(\mathbf{y}_i^t \mathbf{W}_q^{-1} \mathbf{y}_j\right)^2 \\
&\alpha_i = n_i^q \Big/ \left(n_i^q + 1\right); \alpha_j = n_j^q \Big/ \left(n_j^q - 1\right); \quad \mathbf{y}_i = \mathbf{X}_r - \mu_i^q; \quad \mathbf{y}_j = \mathbf{X}_r - \mu_j^q
\end{aligned} \tag{11}$$

If $\Delta_q(r,j,i) \le \rho < 1$ then $|\boldsymbol{W}_{q+1}| < |\boldsymbol{W}_q|$. This condition ensures that the procedure does indeed produce progressively better partitions. Moreover, since $|\boldsymbol{W}_q|$ is bounded by zero, the process must converge in a finite number of steps. (Obviously it is not the convergence itself, but the rate of convergence that justifies this method in practice). A threshold lower than one (*e.g.* $\rho=0.9999$) prevents cycling divergence (that is, catastrophic recurrence of partitions which were abandoned at an earlier stage) due to numerical problems; additionally, it may help to regulate the running time of the algorithm.

The change in the scatter matrix, its inverse, centroids and cardinalities is easily computed from the following relations

$$\begin{aligned}
&\mathbf{W}_{q+1} = \mathbf{W}_q - \alpha_j \mathbf{y}_j \mathbf{y}_j^t + \alpha_i \mathbf{y}_i \mathbf{y}_i^t; \quad 1 - \alpha_j \mathbf{y}_j^t \mathbf{Z}_q^{-1} \mathbf{y}_j \neq 0; \; 1 + \alpha_i y_i^t \mathbf{W}_q^{-1} \mathbf{y}_i \neq 0 \\
&\mathbf{W}_{q+1}^{-1} = \mathbf{Z}_q^{-1} + \frac{\alpha_j \left(\mathbf{Z}_q^{-1} \mathbf{y}_j\right)\left(\mathbf{Z}_q^{-1} \mathbf{y}_j\right)^t}{1 - \alpha_j \mathbf{y}_j^t \mathbf{Z}_q^{-1} \mathbf{y}_j}; \quad \mathbf{Z}_q^{-1} = \mathbf{W}_q^{-1} - \frac{\alpha_i \left(\mathbf{W}_q^{-1} \mathbf{y}_i\right)\left(\mathbf{W}_q^{-1} \mathbf{y}_i\right)^t}{1 + \alpha_i y_i^t \mathbf{W}_q^{-1} \mathbf{y}_i} \\
&\mu_i^{q+1} = \frac{n_i^q \mu_i^q + \mathbf{X}_r}{n_i^q + 1}; \quad \mu_j^q = \frac{n_j^q \mu_j^q - \mathbf{X}_r}{n_j^q - 1}; \; n_i^{q+1} = n_i^q + 1; \; n_j^{q+1} = n_j^q - 1
\end{aligned} \tag{12}$$

To avoid the accumulation of rounding errors, the quantities are computed directly from the data after a number $\nu$ of transfers depending on the data set. ***DetClus*** uses $\nu=200\sqrt{n}*m$.

The sequence of the entities within the data set may exert a profound influence on k-means. An algorithm is said to be combinatorial (MacQueen, 1967) if the criterion, centroids, cardinalities and within-group scatter matrices are updated immediately after a move has been executed in order to take account of the new situation. As a result, the trajectory of the iterative process is dependent, to some extent, on the sequence in which entities are processed and different orderings may yield different clusterings. This problem can be mitigated by randomizing the choice of the entities to be reallocate or by applying data reordering techniques.

In a noncombinatorial k-means algorithm (Forgy, 1965) the moves are executed in parallel in the sense that the entities do not actually change to their new cluster membership until destinations for all entities have been determined. Hence, not only the calculations are substantially simplified, but the iterative process does not suffer from ordering effects. However, unless certain conditions are satisfied (Selim and Ismail, 1984), there is no guarantee of a net improvement in $L(\gamma)$ and no guarantee that the k-means process converges.

A relocation of the entity $\boldsymbol{X}_r$ from cluster $i$ to cluster $j$ causes consequential changes to the centroids $\mu_i$ and $\mu_j$: the former is pulled toward $\boldsymbol{X}_r$ and the latter is pushed away from it. This causes the distances from the centroid of other entities in clusters $i$ and $j$ to decrease, such that the criterion is decreased. If $\boldsymbol{X}_r$ is shifted to its nearest cluster centroid but the centroids are not upgraded the combined effects of several moves like this may actually increase the criterion or, worse, the same reallocations are reproposed in two or more successive steps and no further improvement may be obtained by the algorithm.

Another drawback of the Forgy step is that during the relocation phase it is possible that all entities of a cluster are assigned to other clusters and at the same time no other entity is assigned to the centroid of this cluster. In this way the procedure ends up with an empty clusters and the partition is discarded.

In spite of this potential weaknesses, the Forgy approach can generate fast and reliable k-means algorithms which, nonetheless, tend to be less efficient than algorithms implementing the McQueen approach (Anderberg, 1973, p. 166). On the other hand, it has been experimentally observed that algorithms based on a combinatorial scheme are more susceptible of being trapped in local minima. At present, the effect of the choice combinatorial/non combinatorial reassignment of the entities for the criterion (9) has not yet fully established.

*Option 1: first improving*

The simplest reassigning pass merely consists of scanning -in a random or systematic order- the data set and computing $\Delta_q(r,j,i)$ for $i=1,2,...,k,\ i\neq j$; $r=1,2,...,n$. where the order in which the cluster are tried can also be sequential or random. If $\Delta_q(r,j,i)\leq\rho$ then $\boldsymbol{X}_r$ is immediately reclassified from its present cluster $j$ to cluster $i$ without checking to see if some other transfer would be better. The $n$ entities are then checked in turn to see if another transfer decreases the criterion. For each entity, ***DetClus*** examines at most $(k-1)$ partitions (neighborhood set) derived from the current partition by moving an entity from one cluster to another. It should be noted that when the starting partition is inadequate the "quick" transfers can be slower than more complex searches executed under the other options.

The results of TFI may depend on the sequence in which the entities are processed. If the data sets are formed by compact and isolated clusters, there is a high chance that any arrangement of the data may lead to a global minimum (MacQueen, 1967). Nevertheless, more consistent and reliable comparisons can be performed if the way the entities are selected for the updating phase does not interfere with the minimization process. Peña *et al.* (1999) suggested trying many runs with different arrangements to marginalize out ordering effects, but the number of repetitions deserves further exploration. Fisher *et al*. (1992) argued that arrangements so that consecutive entities are dissimilar lead to good clustering. Further work remains to be done on connections between sorting strategies of the data and recovery rate of combinatorial k-means algorithm based on the determinant. Normally, the transfers are executed in the order in which they appear in the data set, but the flow can be altered by the user. In fact, to reduce the impact of the entity order ***DetClus***, allow randomizing both the choice of the entity to be considered for a move and the choice of the destination cluster. The current configuration of the entities is obtained by shuffling the set of entities. Let $\gamma=(\gamma_1, \gamma_2, ..., \gamma_n)$ be a vector of integers between 1 and $n$. By using the technique suggested by Knuth (1981, p. 139) random permutations of the $\gamma_r$'s is considered. In the same way, the current sequence of the destination clusters is determined by shuffling a vector of integer between 1 and $k$. To alleviate the burden of computations the shuffling of the clusters is performed each five transfers and that of the entities each 20 transfers.

*Option 2: local best-improving*

A first-improving policy may lead to premature convergence of the k-means process. The transfer algorithm can be more effective if a local search is included between iterations. This motivated the development of several search methods to solve the problem of $Min\{|\boldsymbol{W}(\gamma)|\}$. Rubin (1967) suggested examining the potential effect of switching $\boldsymbol{X}_r$ from the cluster it occupies to each other cluster and finding the value satisfying $Min\{\Delta(r,j,i)|\ \Delta(r,j,i)\leq\rho,\ i=1,2,...,k;\ j\neq i\}$ Thus each entity is transferred (if transferred) to the cluster which maximizes the impact on $|\boldsymbol{W}(\gamma)|$ of

the transfer. The entities can be considered either in natural or in a random sequence. If such transfer exists then the process is moved from the current partition to the best partition among the *(k-1)* partitions belonging to the neighborhood set. When there is more than one entity whose transfer gives the same decrease of the criterion, the gaining cluster is selected by choosing the transfer with the smallest *i* among the competitors. The search is repeated -using either deterministic or stochastic sequences- for each entity of the data set. It is evident that option 2 is more computer demanding than option 1 since the latter is interrupted if also the former is interrupted, but this may continue to evaluate transfers also when the TFI does not.

*Option 3: Best global improving*

***DetClus*** performs a complete scanning of the entities and produces the set of candidate transfers $E= \{\Delta(r_h,j_h,i_h)\leq\rho,\ h=1,2,...,\}$. Then the elements of $E$ are sequenced in ascending order and the corresponding transfer executed (provided that $\Delta(r,j,i)\leq\rho$ after each transfer) starting from the first, but discarding those affecting clusters already involved in a reassignment. When there is more than one entity whose transfer gives the same decrease of the criterion, the gaining cluster is selected by choosing the transfer with the smallest *i* among the competitors. The process is iterated until all entities no longer change their membership. Iterations are also stopped if $|\boldsymbol{W}| < 10^{-20}$ to avoid looping and overflowing. Clearly, global strategies are expected to give better results than local ones since an improvement of the local search do not necessarily mean an improvement of the total k-means algorithm. However, global strategies may be questionable under the request of computer resources. In fact, for each pass through the data set ***DetClus*** moves at most *[k/2]* entities which could seem unsatisfactory compared with the number of potential relocations considered by a local search. It should be pointed out, though, that after some initial iterations characterized by quick refinements, local searches tend to settle into sequences of very few and often ineffective moves even when the process is not in the vicinity of a minimum partition. In addition, the results of ***DetClus*** are invariant with respect of the entity order (except when multiple equivalent solution exist), whereas the final solution of combinatorial algorithms incorporating local searches may feel the impact of order dependency.

To avoid array overflow errors the number of transfers between cluster *i* and *j* to be retained should have a fixed upper bound because the number of potential moves (its maximum is *(k-1)n* ) could be greater then the available temporary storage). ***DetClus*** considers a maximum of 1'124'250 moves. The option of retaining only the best transfer for each entity although parsimonious in terms of memory storage and execution time, has been proved much less efficient than considering all the transfers (allowed by the memory size of the program).

.

**2.4 Swapping entities**

Banfield and Bassil (1977) proposed that the interchange of cluster membership between entities is a useful tool for reassigning entities. ***DetClus*** offers the opportunity of a mixed scheme alternating transfers with swaps.

Consider the swap of entity $\boldsymbol{X}_r$ with $\gamma_r = i$ and entity $\boldsymbol{X}_s$ with $\gamma_s = j$, $i \neq j$. The effect on the dispersion matrix is

$$\mathbf{W}_{q+1} = \mathbf{W}_q - \beta(\mathbf{X}_r - \mathbf{X}_s)(\mathbf{X}_r - \mathbf{X}_s)^t + (\mathbf{X}_r - \mathbf{X}_s)(\mu_j - \mu_i)^t + (\mu_j - \mu_i)(\mathbf{X}_r - \mathbf{X}_s)^t \qquad (13)$$

where $\beta = (n_i + n_j)/(n_i n_j)$. The inverse of (13) and its determinant can be computed by repeated applications of the Sherman-Morrison formula.

$$\begin{aligned} &|\boldsymbol{W}_{q+1}| = |\boldsymbol{W}_q| * \Delta(r,s,j,i) = |\boldsymbol{W}_q| * \left(1 - \beta \boldsymbol{f}^t \boldsymbol{W}_q^{-1} \boldsymbol{f}\right)\left(1 + \boldsymbol{g}^t \boldsymbol{B}_q^{-1} \boldsymbol{f}\right)\left(1 + \boldsymbol{f}^t \boldsymbol{A}_q^{-1} \boldsymbol{g}\right) \\ &\boldsymbol{f} = (\mathbf{X}_r - \mathbf{X}_s), \ \boldsymbol{g} = \left(\mu_j^{\mathrm{q}} - \mu_i^{\mathrm{q}}\right) \end{aligned} \qquad (14)$$

$$\boldsymbol{B}_q^{-1} = \boldsymbol{W}_q^{-1} + \frac{\beta\left(\boldsymbol{W}_q^{-1}\boldsymbol{f}\right)\left(\boldsymbol{W}_q^{-1}\boldsymbol{f}\right)^t}{1 - \beta \boldsymbol{f}^t \boldsymbol{W}_q^{-1} \boldsymbol{f}}; \ \boldsymbol{A}_q^{-1} = \boldsymbol{B}_q^{-1} - \frac{\left(\boldsymbol{B}_q^{-1}\boldsymbol{f}\right)\left(\boldsymbol{B}_q^{-1}\boldsymbol{g}\right)^t}{1 + \boldsymbol{g}^t \boldsymbol{B}_q^{-1} \boldsymbol{f}};$$

$$\boldsymbol{W}_{q+1}^{-1} = \boldsymbol{A}_q^{-1} - \frac{\left(\boldsymbol{A}_q^{-1}\boldsymbol{g}\right)\left(\boldsymbol{A}_q^{-1}\boldsymbol{f}\right)^t}{1 + \boldsymbol{f}^t \boldsymbol{A}_q^{-1} \boldsymbol{g}}; \ \mu_i^{q+1} = \mu_i^q + n_i^{-1}\boldsymbol{f}; \ \mu_j^{q+1} = \mu_j^q - n_j^{-1}\boldsymbol{f} \qquad (15)$$

For each scan of all possible interchanges between different clusters ***DetClus*** implements the swaps (if any) which most reduce the criterion, provided that no cluster is involved in more than one swap and that $\Delta(r,s,j,i) \leq \rho$ after each swap. This condition ensures that the procedure does indeed produce progressively better partitions. Since the criterion $Min\{\boldsymbol{W}(\gamma)\}$ corresponds to a sum of squares, the process of relocating only those entities which yield a reduction must converge because a sum of squares cannot be indefinitely reduced.

As was previously noted, the k-means algorithm can be interrupted after the first improvement found in the neighborhood set or after examining the whole neighborhood set. In the former case, a maximum of $n(n-1)/2$ candidate partitions are evaluated while, in the latter, exactly $n(n-1)/2$ alternatives must be analyzed. In the first case, an order dependency may be introduced which can be ameliorated randomizing the choice of the pairs to be swapped.

*Option 1: Post processing stage*

With this option the user activates a hybrid process oscillating between the transfers stage and the swaps stage. The whole data set is reprocessed until there is no further improvement in the quality of the clustering by means of a transfer; only then the swaps stage is executed repeatedly for all pairs of entities until a new convergence occurs. If one or more swaps are found beneficial, then the transfers stage is restarted. The iterations continue until the membership of the clusters stop changing. The hybrid scheme denoted as option 1 should be essentially considered as a way of overcoming a local minimum. The swopping, is a heuristic technique in the sense that its failure to produce a better solution does not mean that the actual partition is the best. However, it reinforces our confidence in it.

*Option 2,3. Mixed strategies: transfer+swaps*

The transfers are applied for the first pass across all entities then the swaps for the second, and proceed in this fashion until a minimum of the criterion is reached. Banfield and Bassil (1977) considered a single search of the *n(n-1)/2* pairs of entities although further repetitions (after recomputing the centroids) could led to better partitions.

*Option 4,5. mixed strategies: swaps+transfers*

In this case the swaps are used for the first stage then transfers for the second and continue oscillating until there are no entities that change their cluster membership. The mixed strategies should help in applying k-means with inadequate starting partitions.

The swapping pass (options 2-4) can be combined with the transferring pass (option 1-3) generating 12 mixed schemes: TFI+SFI, TFI+SGBI, TBLI+SFI, TLBI+SGBI, TGBI+SFI, TGBI+SGBI, SFI+TFI, SGBI+TFI, SFI+TLBI, SGBI+TLBI, SFI+TGBI, SGBI+TGBI. The pure schemes: TFI, TLBI, TGBI reprocess the whole data set and terminates when there are no entities that change their cluster membership. The mixed schemes have two distinct alternating strategies: either the transfers are applied for the first pass across all entities then the swaps for the second, and proceed in this fashion until a minimum of the criterion is reached or the swaps are used for the first stage then transfers for the second and continue oscillating until convergence occurs. In both cases, mixed schemes suffer from ordering effects, with the exception of TGBI+SGBI and SGBI+TGBI.

**2.5 Simulation results**

Tarsitano (2002) has analyzed and compare 17 different relocation methods for the k-means algorithm implementing the Friedman-Rubin criterion (given that the number of natural clusters is known and the order of entities within the data set is fixed).

The key findings are listed below.

1. The scheme TGBI, unexpectedly, scores top marks in terms of convergence speed significantly better than any other scheme. In this sense, it is a natural candidate for clustering large data sets, at least for applications where a reasonably good initial classification is available.

2. The mixed schemes are uniformly less rapid than pure schemes and the difference between execution times reaches a maximum -as it should be suspected- when the globally best transfer is coupled with the globally best swap. On the other hand, when swaps are performed, TGBI+ are faster than TLFI+ which are, in turn, faster than TFI+. The same ranking is found for the tandems lead by SFI and for those lead by SGBI. The durations of TGBI+ and SGBI+ are higher than any other mixed scheme by orders of magnitude. It is evident that the swapping stage is a time-consuming task because it compares an entity with the entire data set. Worth of note is that the results of combinations S+T compares favorably with those of the reverse combinations T+S. Banfield and Bassil (1977) have ignored mixed methods of the type S+T which, on the contrary, seems to generate efficient schemes.

3. Coleman *et al.* (1999) argue that a TFI strategy seems to be preferred to a TLBI strategy for the problem of classification to minimize the determinant criterion. Ismail and Kamel (1984) indicate that TLBI is more susceptible to being trapped at a local minimum than TFI, at least for algorithm guided by $Min\{Tr(\boldsymbol{W}(\gamma))\}$. On the other hand, Zhang and Boyle (1991) found that TFI and TLBI are indistinguishable. These findings were not confirmed by my experiments. For k-means algorithms based on $Min\{|\boldsymbol{W}(\gamma)|\}$, TGBI outperforms all the other methods, regardless the number of variables, the number of clusters and the structure of the cardinalities. The inclusion of a global search determining a chain of reassignments each of which is the best taken from among the available reassignments is generally beneficial for improving both the rate of convergence and the accuracy of the final partition. Moreover, TGBI is indifferent to the order of data whose influence on other schemes is complex and unpredictable. For medium sized data sets the algorithm runs quite efficiently. Huge data sets are precluded because the large values of $nm$ would require excessive computer resources.

4. The mixed schemes T+ obtain (but non always) some refinement of the final partition over the respective pure schemes. Similar results are found for SFI+ and SGBI+. Nonetheless the impact of the swaps over the quality of the solution is limited and the time needed for each convergence may not be worth the extra computation. Mixed schemes are more likely to work better for poor starting conditions, but the limited impact on the classification adequacy does not compensate the extra energy expended for these procedures. The experiments indicate that com-

binations of different strategies may provide significantly less good performance than do their isolate application. In particular, the swaps, not only are time consuming, but also tends to block the process after very few iterations. In practice, the swaps should be essentially considered as a way of getting out of a local minimum.

5. Schemes of the type S+ tend to yield better solutions in terms of stability and accuracy than T+. Most likely the phenomenon is due to the major ability of the swaps to use more productively the fact that most of the changes in cluster membership occurs at the first few iterations (Anderberg, 1973, p. 163)

6. For data sets divided into even clusters, the recovery rate is steadily higher than for disparate sized clusters and the differences becomes more pronounced as the number of clusters increases. This is aligned with the conjecture that $Min\{|W(\gamma)|\}$ encourages the formation of partitions with clusters of equal size if the separation between the clusters is not large (Scott and Symons, 1971; Everitt *et al*. 2001, p. 94).

7. An interesting point is that the dimension of the problems and the number of clusters did not affect the convergence of either of the algorithms implemented in **DetClus.**.

*Overcoming local minima*

The problem of IPM's is that the local minimum $\gamma^*$ may not be the global minimum. Rubin (1967) remarked that two type of problems cause local minima:

1) Two homogeneous but unrelated clusters are united while other clusters may be well formed;
2) The centres of the clusters do not allow a very stable classification of hybrid entities.

The first situation affers directly with the problem of the number of cluster and will be discussed in section 4. **DetClus.** attempts to circumvent local minima due to the second situation by swaps. The swopping, as many other techniques for overcoming a local minimum, is heuristic in the sense that its failure to produce a better solution does not mean that the actual partition is the best. However, it reinforces our confidence in it. It must be said that the swopping phase, rarely provides an improvement and may be ignored if the algorithm starts from a good configuration.

*Limitations*

An inherent limitation of the k-means algorithms included in **DetClus** is that their final configuration does not necessarily coincide with one of the desired global minima. Since all the schemes do only descent moves, they are not able to force the process out of the current valley and eventually fall into a deeper one. The development of mixed algorithms which combine the best elements of the transfers/swaps with a non descent technique would be a significant contribu-

tion. Additional work is needed to determine the most appropriate strategy of alternating transfers and swaps and to keep the algorithms from taking too many iterations in regions where insufficient progress is being made. For instance, hybrid processes which oscillate between a trasferring stage which reprocess the whole data set until there is no further improvement in the quality of the clustering and only then a swapping stage is used repeatedly for all pairs of entities, should help in applying k-means with inadequate starting partitions. Furthermore, two-phase strategies in which a first-improving transfer pass is applied if the entity number is odd otherwise a best-improving pass is executed (or *vice versa*) can be considered (either as isolate application or combined with swaps) to devise a better k-means algorithm. Moreover, strategies in which the swopping phase is done periodically or randomly could be devised.

The performances of iterative partitioning methods are mainly affected by the intensity of the clustering. The procedures described in this section are all appropriate when the clusters form essentially compact clouds that are fairly well separated from one another. If the clusters are close to one another (even by outliers or hybrids), or if their shapes are not hyper-ellipsoidal, the results of clustering can vary quite dramatically. In fact for poorly defined clusters the misclassification rate reaches unacceptable levels even if the method is valid and consistent with the data-generating process. Furthermore, as Mineo (1986) pointed out, it is more difficult to determine a good starting point and, as a consequence, the algorithm is more likely to stop on local minima

## 3. Initialization methods

The k-means algorithms described in the previous section converge finitely to a partition $\gamma^*$ that is locally minimal for $|\boldsymbol{W}(\gamma)|$. The convergence is deterministic given the initial configuration, but the quality of the minimum is not guaranteed. The efficacy of a k-means algorithm is influenced by many factors. Most obvious is the starting partition. In fact, k-means algorithms have differential recovery rates depending on the quality of the starting configuration. So far no attempt has been made to set up a procedure that works well on every occasion.

The reason for this is simply that what is most appropriate for one data set may not be so for another and, unfortunately, there is no simple, universally good solution to this problem (Duda et al, 2001, p. 550). In some case it is possible to obtain excellent results by taking the first $k$ entities as typical representatives; in others, only sophisticated and computationally expensive methods may provide an initial partition acceptably close to the final solution.

Furthermore, the concept of "best" is a compromise between accuracy and computation cost which, for this reason, cannot lead to an initialization method that outperforms all the others on all the data sets. Many available procedures invite the user to have a hieratic confidence in a built-in initialization method (no further details) which regularly finds good clusters provided that they exist and the user gives the correct number of clusters to detect. Such a guarantee cannot be given.

As the cluster analysis has evolved, a a wide variety of techniques has emerged for choosing the first $k$ centroids (or, alternatively, for specifying an appropriate starting partition $\gamma^{(0)}$). Anderberg (1973), Hartigan (1975 Blashfield *et al*. (1982), and Peña *et al*. give a brief summary of a number of different procedures by which an iterative partitioning could be triggered and more can be found (*e.g*. Mineo,1985, Al-Daoud and Roberts,1996).

If an inadequate initialization is performed two puzzling phenomena tend to appear. First, the algorithm may be interrupted at a lower value of the criterion not corresponding to a greater recovery rate (Coleman *et al.*, 1999). Second, the minimum partition found may not be unique as other partitions may give the same criterion value which, in addition, may be associated with a different degree of clustering effectiveness. There is little chance to avoid these problems because the surface defined by $|\boldsymbol{W}(\gamma^{0})|$ is usually flat and contains many local minima.

Repetition of the procedure with different partitions appear to be a reasonable method to face this problem. Moreover, it can give good indication of the sensitivity of the final solution $\gamma^{*}$ to the starting partition. It should be emphasized, though, that $|\boldsymbol{W}(\gamma^{0})| < |\boldsymbol{W}(\delta^{0})|$ does not necessarily imply that $|\boldsymbol{W}(\gamma^{*})| < |\boldsymbol{W}(\delta^{*})|$. Therefore, a user is advised to try several initialization methods on a given data set. In fact, **DetClus** uses each starting partition as a separate basis for the subsequent phases and the classification vector corresponding to the lowest value of $|\boldsymbol{W}(\gamma)|$ is chosen as final clustering. It hardly need adding that the search of the initial configuration takes very much longer then the entire algorithm (the problem is even serious when $n$, $k$, and $m$ are large). However, the advantages in terms of partitional adequacy of the final solution far outweigh the consumption of computer time. Of course, multiple restarts may be incompatible with large data sets, at least for the actual technology of the combination hardware/software.

***DetClus*** determines $\gamma^{0}$ by trying several effective techniques which can be classified in two categories: deterministic and random.

### 3.1 Deterministic techniques.

The deterministic techniques yield an initial partition which is unique in that it is found optimizing a suitable objective function. The partition for which ***DetClus*** obtains the best results is written in the output file as optimal solution for the given value of $k$.

### 3.1.1 Best among naive methods

This command calls three different procedures characterized by rapid movements of the entities and quick computations. To avoid generating unfeasible partitions all clusters with no entities assigned to them receive a randomly chosen entity from the largest cluster.

*Option 1: mean entity*

Hartigan (1975, p. 88) proposed a quick initial clustering based on the simple arithmetic mean of the variables for each entity. ***DetClus*** uses a weighted average of the variables which standardizes the variables by their sample variances. In particular, the observed value of the $m$ variables for the r-th entity is summarized by

$$S_r = \sum_{j=1}^{m} w_j x_{r,j}; \qquad w_j = \frac{\sigma_j^2}{\sum_{i=1}^{m} \sigma_i^2}, \; j = 1,2,\ldots,m \tag{16}$$

where $\sigma_j^2$ is the sample variance of $X_j$. The r-th entity is assigned to the cluster $C_i$ if

$$\left[(k-1)\left(\frac{S_r - S_{min}}{S_{max} - S_{min}}\right) + 1\right] = i = \gamma_r, \quad r = 1,2,\ldots,n \tag{17}$$

*Option 2: leading component*

Hartigan (1975, p.102). Let $w_j$ for $j=1,2,\ldots,m$ be the factor loadings of the first principal component of $(n-1)^{-1}\boldsymbol{T}$ and let

$$S_r = \sum_{j=1}^{m} w_j x_{rj}, \quad r = 1,2,\ldots,n \tag{18}$$

The cluster membership is given by (17). Of course, if the variables have been expressed as factor score, the rule (17) applies to the first variable of the transformed data set. It must be said that the averaging features applied by option 1 and 2 could destroy information contained in the multivariate data.

*Option 3: quantiles*

The ordered scores $S_r$, $r=1,2,\ldots,n$ of the dominant factor of $(n-1)^{-1}\boldsymbol{T}$ are divided into $k$ slices with approximately the same number of entities. The membership of the entities is determined according to the rule

$$\gamma_r = j, \; if \;\; S_{\left(\frac{j-1}{k}\right)} \le S_r \le S_{\left(\frac{j}{k}\right)}; \; S_{0.0} = S_{min}, \;\; S_{1.0} = S_{max} \tag{19}$$

where $i = [nt + 0.5]$; $\alpha = nt + 0.5 - i$

### 3.1.2 Best among built-in techniques (simple methods)

These methods are considered "simple" because they perform a single pass through the data set, but require an estimate $\boldsymbol{W}^*$ of $\boldsymbol{W}$. Since the cluster membership is unknown before the analysis some approximate procedure must be used (see section 3.1.3)

*Option 1: Sequential splitting*

Let $g$ be the number of clusters already formed and let $h$ and $i$ indicate, respectively, the cluster and the variable where the coefficient

$$CD = \frac{2\sqrt{\dfrac{\sum_{\gamma_r = h}\left[x_{ri} - \mu_{ih}\right]^2}{n_h}}}{\left[|L_{ih}| + |U_{ih}|\right]};\quad L_{ih} = \underset{\gamma_r = h}{Min}\left\{x_{ri}, i = 1,\ldots,n_h\right\};\quad U_{ih} = \underset{\gamma_r = h}{Max}\left\{x_{ri}, i = 1,\ldots,n_h\right\} \tag{20}$$

is higher. Index (20) increases as the relative variability increases and it is immune from standardization bias; moreover, the denominator does not vanish unless $X_{ri} \cong 0$ (in this case $CD = 0$). The denominator of $CD$ is not an average since its value may fall outside the sample range. Suppose that the current number of entities in cluster $C_h$ is $n_h > 1$ and that $X_i$ is not constant in $C_h$. Then cluster $C_h$ can be splitted as follows

$$\gamma_r^0 = \begin{cases} g & if \quad x_{ri} \le M \\ g+1 & if \quad x_{ri} > M \end{cases};\quad \gamma_r = h;\quad M = \underset{1 \le t < n_h}{Max}\left\{\frac{\left[\sum_{i=1}^{t} x_{(t)i}\right]^2}{t} + \frac{\left[\sum_{i=t+1}^{n_h} x_{(t)i}\right]^2}{n-t}\right\} \tag{21}$$

Formula (21) maximizes the between-group sum of squares for the i-th variable (Engelman and Hartigan, 1969; Anderberg, 1973, pp. 45-46 ). The split separates the cluster of points above the mean $\mu_i$ from the cluster of points below the mean. Centroids are then computed for each cluster by averaging coordinates of its members. In practice, a Forgy step though the data is executed, that is the entities do not change to their new cluster membership until all assignment have been evaluated. The assignment of the entities to the clusters is based on Mahalanobis distance with metric $(\boldsymbol{W}^*)^{-1}$. At the end of step the cluster centroids are updated to be the averages of entities contained within them. No further iterations are performed. Splitting continue until $g+1=k$. A clusters that has less than one entities as its members id discarded.

*Option 2: ordered distance from the total mean*

This procedure was proposed by Hartigan and Wong (1979). The entities are first sorted by their distances to the overall mean vector $\mu$ of the data set; then, the cluster centroids $P_j$, $j=1,2,\ldots,k$,

are chosen to be the entities labelled *1+(j-1)b, j=1,2,...,k* with *b=[n/k]*. The classification vector $\gamma$ is derived according to

$$\gamma_r = j \;\; if \;\; \left(X_r - P_j\right)' W^{*-1} \left(X_r - P_j\right) \le \left\{\left(X_r - P_i\right)' W^{*-1} \left(X_r - P_i\right)\right\};\; i = 1,2,\ldots,g \tag{22}$$

Formula (22) implies that an entity which need to be assigned to one of the clusters is identified with the cluster to which it is closest as judged by the Mahalanobis distance based on the metric $W^*$. If an entity is at the same distance from several centroids it is by convention assigned to the cluster $C_j$ with the smallest index *j* among the competitors.

*Option 3-6: farthest neighbor*

These procedures consist of 3 steps.

Step_1. Determine the first centroid $P_1$. Two alternatives may be considered:

a) the first centroid is the entity which is nearest to $\mu$, the mean vector of the data set.

$$P_1 = X_s \Rightarrow \left(X_s - \mu\right)' W^{*-1} \left(X_s - \mu\right) \le \left(X_r - \mu\right)' W^{*-1} \left(X_r - \mu\right) \;\; r = 1,2,\ldots,n \tag{23}$$

b) the first centroid is the entity which has the greatest distance from $\mu$.

$$P_1 = X_s \Rightarrow \left(X_s - \mu\right)' W^{*-1} \left(X_s - \mu\right) \ge \left(X_r - \mu\right)' W^{*-1} \left(X_r - \mu\right) \;\; r = 1,2,\ldots,n \tag{24}$$

Step_2. Let $P=\{P_1, P_2, ..., P_g\}$ be the current set of centroids. The (g+1)-st centroid may be chosen according two alternative rules

$$P_{g+1} = X_s \Rightarrow \underset{P_j \in P}{Min} \left(X_s - P_j\right)' W^{*-1} \left(X_s - P_j\right) \ge \underset{P_j \in P}{Min} \left(X_r - P_j\right)' W^{*-1} \left(X_r - P_j\right) \tag{25}$$

$$P_{g+1} = X_s \Rightarrow \underset{P_j \in P}{Min} \sum_{j=1}^{g} \left(X_s - P_j\right)' W^{*-1} \left(X_s - P_j\right) \ge \underset{P_j \in P}{Min} \sum_{j=1}^{g} \left(X_r - P_j\right)' W^{*-1} \left(X_r - P_j\right) \tag{26}$$

Step_3. Execute a Forgy steps through the data set *i.e.* the changes caused by each entities are accumulated and executed at the end of the cycle. The classification vector is determined by assigning all entities to the most similar centroid. Only one complete pass through all the entities is executed. The assignment of the entities to the clusters with the nearest centroid is based on (22). Step_2 and Step_3 are repeated until *k* centroids have been selected.

Methods implementing the farthest-neighbor policy have the advantage of ensuring that extreme entities appear in the initial configuration, but have the drawback of including as centroids atypical entities such as outliers which are unduly emphasized by the Mahalanobis norm.

***DetClus*** uses the following labels:

*Option_3: (total mean, maximum distance)* *Option_4: (total mean, mean distance)*
*Option_5: (farthest entity, maximum distance)* *Option_6: (farthest entity, mean distance).*

### 3.1.3 Preliminary estimation of the within-clusters matrix

Art *et al*. (1982) proposed an algorithm to compute an estimate of **W** without knowing the cluster structure but assuming that the clusters have different means and a common covariance matrix. The standard multivariate analysis decomposition (10) can also be made in terms of pairwise differences:

$$\frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)^t=\underbrace{\frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)^t}_{Within}+\underbrace{\frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)^t}_{Between} \tag{27}$$

$$=W^*+B^*$$

The first term on the right side of (27) involves all the pairs which belongs to the same clusters, and the second term involves all the distance measurement occurring between those pairs where one entity comes from cluster *i* and the other entity comes from cluster *j* with $i \neq j$. No explicit indication is made to the classification vector. The left sides of (10) and (28) are equal. Therefore $\mathbf{T}=\mathbf{W}^*+\mathbf{B}^*$. Under normal sampling assumptions, with $X_{ij} \sim N(\mu_i, \Omega)$ the expected values of $\boldsymbol{W}$ and $\boldsymbol{W}^*$ are

$$E(\boldsymbol{W})=(n-k)\Omega, \quad E\left(\boldsymbol{W}^*\right)=\left(\frac{\sum_{i=1}^{k} n_i^2-n}{n}\right)\Omega \tag{28}$$

Hence $\boldsymbol{W}$ and $\boldsymbol{W}^*$ can be used to construct an unbiased estimate for $\Omega$, but $\boldsymbol{W}^*$ gives relatively more weight to large clusters than does $\boldsymbol{W}$. Naturally, since the cluster structure is unknown neither $\boldsymbol{W}$ nor $\boldsymbol{W}^*$ can be computed. The initialization of an iterative partitioning, however, requires something of less stringent and even a a crude estimate of $\Omega$ can be very helpful. Generalizing the idea of Art et al (1982) a first approximation to $\boldsymbol{W}^*$ can be obtained by

$$\boldsymbol{W}_{(1)}^*=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_h\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)^t; \quad h=j+(i-1)n-\frac{i(i+1)}{2} \tag{29}$$

$$with \quad \delta_h = \begin{cases} f\left[\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)^t \boldsymbol{M}^{-1}\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)\right] & if \quad h \le q \\ 0 & if \quad h > q \end{cases}; \quad \delta_h \ge 0, \ \sum_{h=1}^{q} \delta_h = 1 \tag{30}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are among the closest $q$ pairs in terms of the metric $\boldsymbol{M}$. The weight function is such that $f'(x)<0$ *for* $x>0$, that is, the weight $d_b$ is a non increasing function of the distance between the pair $(i,j)$: the larger the distance is the smaller is the weight attached to the pair. Then (n-1)/2 possible pairs of entities need not to be sorted as long as it can be established that $h>q$ or not. The integer $q$ is chosen conservatively small to avoid contamination by between-cluster pairs.

Next $\boldsymbol{W}^*_{(2)}$ is formed in the same manner except that a new squared generalized distance is used to define the coefficients $d_h$, that is $\boldsymbol{M}=\boldsymbol{W}^*_{(1)}$

$$\delta_h = f\left[\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)^t \left[\boldsymbol{W}^*_{(1)}\right]^{-1}\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)\right] \tag{31}$$

The algorithm continues in a like manner until the process stabilizes, which it usually does rather quickly.

The estimation procedure is controlled by the following parameters:

1) The first metric. Art *et al.* (1982) used $\boldsymbol{M}=\boldsymbol{I}$, that is the first allocation is made by using Euclidean distance, although $\boldsymbol{M}=\boldsymbol{T}$ seems a more plausible choice when the data consist of a number of variables measured in different scales and $\boldsymbol{T}$ is well-conditioned. Using the total covariance matrix as the first estimate, while simple and obvious, also ends up ignoring possible clusters in the data. Another plausible choice is $\boldsymbol{M}=\boldsymbol{V}=diag(v_1, v_2, ..., v_m)$. It should be noted that choosing a diagonal is justified only when the variables are uncorrelated or weakly correlated. If this fact is not taken into account, the measure of closeness of the entities suffers.

2) The number of pairs. Art *et al.* (1982) and Gnanadesikan *et al.* (1993) suggested $q=(n/3)(n/k-1)$ neglecting the number of dimensions. More reasonable values can be found in the range $4[m^2+m(2k-1)] \le q \le (n-k)(n-k-1)/3$.

3) The weights. Art *et al. used* $d_h=1/q$ which have, undoubtedly, some advantages from a computational point of view. In fact, the sorting of the distances is not necessary because it is sufficient to determine the smallest $q$ distance, and these may be unsorted if the scope is their unweighted sum. However, the partial sorting involved in this approach has an high cost in term of storage and the gain in execution time is irrelevant. Moreover, the heapsort suggested by Art *et al*. has a mean time which is inferior to the recursive quicksort implemented by **DetClus**.

After some experiments the following formula has given better results

$$\delta_h = \frac{\alpha(1-\alpha)^h}{1-(1-\alpha)^{q+1}}, \ h = 1,2,\ldots,q, \ 0 < \alpha < 1$$

where *h* indicates the h-th closest pairs.

4) The measure of closeness. Art *et al.* defined

$$\varepsilon_{i+1} = trace\left[\left(\boldsymbol{W}^*_{(i)}\left[\boldsymbol{W}^*_{(i+1)}\right]^{-1} - \boldsymbol{I}\right)\left(\boldsymbol{W}^*_{(i)}\left[\boldsymbol{W}^*_{(i+1)}\right]^{-1} - \boldsymbol{I}\right)\right] \tag{32}$$

and convergence is considered satisfactory if $\varepsilon_{i+1} \leq \varepsilon$. An alternative measure is the Jeffreys divergence

$$\varepsilon_{i+1} = \frac{1}{2} trace\left[\left(\boldsymbol{W}^*_{(i)}\left[\boldsymbol{W}^*_{(i+1)}\right]^{-1} - \boldsymbol{I}\right)\left(\left[\boldsymbol{W}^*_{(i)}\right]^{-1}\boldsymbol{W}^*_{(i+1)} - \boldsymbol{I}\right)\right] \tag{33}$$

which allow us to measure the distance between two hypothesis $\boldsymbol{W}=\boldsymbol{W}_i$ *vs* $\boldsymbol{W}=\boldsymbol{W}_{i+1}$ in the case of a multidimensional sample stemming from one of two schemes relative to normal distribution in $R^m$. However, (32) is less computer demanding than (33).

***DetClus*** implements the procedure with **M=I**, $\delta_h = 1/q$ , $q = min\{5[m^2 + m(2k_2 - 1)], n(n-1)/2\}$and (32) with $\varepsilon = 0.001$. Iterations are also stopped after 30 iterations.

The fact that W* needs a multiplicative constant to make it an unbiased estimator of W is not relevant since a clustering based on $\boldsymbol{W}^*$ is invariant with respect of the transformation $\boldsymbol{W}^* = a\boldsymbol{W}^+$ with $a > 0$. The main drawback of this procedure is that becomes inapplicable for large data sets both for the storage and for the sorting of the distances. For $n > 1500$ ***DetClus*** chooses the closest $q$ pairs in a random sample (with replacement) of pairs of size 1'124'250. The weights are given by (32) with $\alpha = 0.0001$. The procedure is stopped after ten iterations or if $\varepsilon_{i+1} \leq \varepsilon$.

A similar method for obtaining an estimate of W is available in the Acelus procedure implemented in the SAS procedure **Fastclus.** However, if the population clusters have very different covariances matrices the procedures outlined above is of no avail.

### 3.1.4 Best among built-in techniques (elaborate methods)

Under this command are comprised four procedures which are computational expensive in that consider, repeatedly, the Mahalanobis distance among all the pairs of entities. None of them is practical when it comes to solving large problems as all of them can become prohibitively expensive even with present-day high-speed computers solution. Moreover, some of them require a large amount of space for storage purposes.

*Option_1: complete link centroids*

Kennard and Stone, (1969) proposed a sequential method to select initial centroids having as even a spread as possible over the variable space. The first two tentative centroids are selected by choosing the two entities that are farthest apart

$$P_1 = X_s,\ P_2 = X_t \Rightarrow \left(X_s - X_t\right)' W^{*-1}\left(X_r - X_w\right) \geq \left(X_r - X_w\right)' W^{*-1}\left(X_r - X_w\right) \tag{34}$$

The entities are assigned to the nearest cluster according to (22). Then a Forgy pass is applied for the reassignment of all entities until all entities. Let $P=\{P_1, P_2, ..., P_g\}$ be the current set of centroids. The (g+1)-st leader is chosen according to (25). The procedures continue until $g=k$.

*Option_2: average link centroids*

Same as option 1, but the (g+1)-st leader may be chosen according to (26). This alternative was suggested by Sadocchi (1977).

*Option_3: representative istances*

Kaufman and Rouseeuw (1990). The first centroid is the most typical member of the data set, that is, the entity $P_1$ such that

$$\sum_{r=1}^{n}\left(X_r - P_1\right)' W^{*-1}\left(X_r - P_1\right) \le \sum_{r=1}^{n}\left(X_r - X_s\right)' W^{*-1}\left(X_r - X_s\right);\ s = 1,\ldots,n \tag{35}$$

Let $P=\{P_1, P_2, ..., P_g\}$ be the set of the current centroids. A new centroid is chosen among the not yet selected entities according to

$$\begin{aligned} &P_{k+1} = X_r \ \Rightarrow \sum_{j=1}^{n} C_{jr} \ge \sum_{j=1}^{n} C_{ji},\ i = 1,2,\ldots,n;\ C_{ji} = Max\left\{D_j - d_{ij};0\right\} \\ &d_{ij} = \left(X_i - X_j\right)' W^{*-1}\left(X_i - X_j\right);\ D_j = \underset{P_s \in P}{Min}\left\{\left(X_j - P_s\right)' W^{*-1}\left(X_i - P_s\right)\right\} \end{aligned} \tag{36}$$

The procedures continues until $g+1=k$. By construction, each cluster has at least one entity. Peña *et al. (1999)* state that (36) chooses as leaders the entities that promise to have around them a higher number of other entities.

*Option_4: divisive analysis (Di.Ana)*

An iterative divisive technique is applied. In practice the Diana algorithm of Kaufman and Rousseeuw (1990, ch. 6) has been extended to rectangular matrices.

The essence of this methods consecutive partition into clusters. Initially, set $C_1=D$. **DetClus** searches for the entity $X_r$ which has the largest average Mahalanobis distance $d(X_r,C_1)$ from all other entities belonging to cluster $C_1$. The entity $X_r$ is discarded from $C_1$ and considered the first entity of the new cluster $C_2$. Let $d(X_s,C_1)$ and $d(X_s,C_2)$ be, respectively, the average distance from the entities in $C_1$ and the average distance in $C_2$. For $X_s$, $s=1,2,\ldots,n$ is left in $C_1$ if $d(X_s,C_1) < d(X_s,C_2)$ otherwise is moved to $C_2$. If $k>2$ then the cluster with the largest diameter (5) is splitted until $k$ clusters have been created.

### 3.2. Random procedures

The random technics generate an initial partition which is independent of the data set. in particular, a pseudorandom sample of ν partitions is considered and the algorithm run for every single partition.

The size $\nu$ is crucial. A larger set of test partitions will make it more likely that $\gamma^{(0)}$ is near to $\gamma^*$, but will also increases the time taken to carry out the search. As an example, the well-known package MIKCA constructed by McRae (1971) starts by analyzing $\nu=3$ different set of randomly chosen leaders; Symons (1981) selected the initial solution from among $\nu=32$ randomly-generated partitions. Casgrain (Le Progiciel R v4.0d6, 2001) has a default value of 100 for ν. Späth (1985, p. 155) criticized heuristic and more elaborate methods for finding a single "good" starting partition and preferred repeating (in his examples, for 20 times) the entire process choosing at random the initial configuration. These values are too small to be really useful. Peña et al. (1999) used $\nu=1000$ initial partitions which is perhaps too large for many data sets. If our objective is to find a partition that is in the top $\alpha$% of $P(n,k)$ and we test a random sample without repetitions of $\nu$ partitions belonging to $P(n,k)$ then the probability of getting such a partition is $p=1-(1-\alpha)^{\nu}$ which implies $\nu=[Ln(1-p)/Ln(1-\alpha)]$. If $\alpha=0.01$ and $p=0.99$ then there is a better than 99% chance that $\nu=458$ will provide a partition which lies in the top 1% of $P(n,k)$. Of course, the top percentile may include highly unsatisfactory partitions.

In a sense $[\sqrt{(nmk)}]$ represents a reasonable compromise between the accuracy of the preliminary search and the duration of a computer run. It hardly need adding that the search of the initial configuration takes very much longer then the entire algorithm (the problem is even serious when $n$, $k$ and $m$ are large). However, the advantages in terms of partitional adequacy of the final solution far outweigh the consumption of computer time.

In **DetClus** the number of partitions to be tried is supplied by the user (the default value is $\left[\sqrt{nmk_2}\right]$).

#### 3.2.1 Random points methods

These method sample the space of the variables determining a centroid as a random point in the convex hull defined by the observed values of the variables.

*Option_1:*

Anderberg (1973, p. 157) suggested the following method to determine the first centroids. Let $L_i$ and $U_i$ represent the minimum and maximum values of the i-th variable for the given data set. Then $R_i=U_i-L_i$ is the sample range of $X_i$.

The coordinates of the leader $P_j$ for the i-th variable are given by

$$m_{ij} = L_i + u_{ij}R_i; \quad i = 1,\ldots,m; \;\; j = 1,\ldots,k \tag{37}$$

where $u_{ij}$ is a uniform random number from *[0,1]*. The starting classification vector is determined according to (22).

*Option_2:*

The total mean of the data set $\mu=(\mu_1,\mu_2,\ldots,\mu_m)$ is chosen as reference point a randomly perturbed to define the centroids of the clusters. More specifically, the coordinates of the k m-dimensional centroids are given by

$$m_{ij} = \begin{cases} \mu_i + u_{ij}\left(U_i - \mu_i\right) & if \;\; e_{ij} < 0.5 \\ \mu_i - u_{ij}\left(\mu_i - L_i\right) & if \;\; e_{ij} \geq 0.5 \end{cases}; \qquad i = 1,\ldots,m; \;\; j = 1,\ldots,k \tag{38}$$

where $u_{ij}$ and $z_{ij}$ are independent uniform random number from (0,1).

The difficult with these schemes is that the resulting centroids are different estimates of the total mean vector and their separateness is questionable. Moreover, unless the data set "fills" the m-dimensional space, some of the centroids may be quite distant from any of the entities and the clusters built around them will have no members. This problem can be attenuated by considering more centroids and eliminating the candidates that are too close. To this end, ***DetClus*** generates *4k* candidates and, among these, selects the best $k$ centroids by applying the Kennard-Stone procedure of section 3.1.2 (*option5)* but ignoring the Forgy step which would be scarcely useful in this context. All clusters with no entities assigned to them receive a randomly chosen entity from the largest cluster

### 3.2.2 Random permutation of representative values

The range of each variable $X_j$, *j=1,2,...,m* is divided into k group. With the i-th group associate a value $m_{ij}$ and imagine that each entity put in the i-th group is given the value $m_{ij}$ for the j-th variable. Then we have a matrix (kxm) of representative values which express the peculiarities of the data set.

Consider a random integer $1 \leq s \leq k^m$ and convert $s$ into the subscript vector $\boldsymbol{I}=(i_1,i_2,\ldots,i_k)$ with $1 \leq i_h \leq k$, *h=1,2,...,k* (see O'Flaherty and MacKenzie,1982).

Then the i-th coordinate of g-th centroid is defined $P_{gj} = m_{i_g,j}$ for *j=1,2,...,m*. Finally, a random

samples without replacement of $k$ vector $\boldsymbol{I}$ from the set of $k^m$ possibilities (cf. Bissell, 1976) is generated to define the $k$ centroids. The initial classification vector $\gamma^0$ is obtained by applying (22). The number of repetitions $\nu$ of this procedure is specified by the user. If $\nu > C(k^m,k)$ then all the combinations are considered as candidate block of centroids.

*Option 1: uniform distributions*

The values of each variable are arranged in ascending order and divided into $k$ blocks. The first $(k-1)$ blocks include $n_j = b = [n/k]$, $j = 1,2,\ldots,k-1$ whereas the remaining $n_k = n-(k-1)b$ entities are allocated to the last block. Suppose that $n_0 = 1$ and let $m_{ij}$ be the partial mean of the block

$$m_{ij} = \frac{\sum_{r=n_{i-1}}^{n_i} x_{(r),j}}{n_i}; \quad i = 1,2,\ldots,k; \quad j = 1,2,\ldots,m \tag{39}$$

*Option 2: partial medians*

It is similar to the first option, but the centroids (31) are replaced by the medians of the blocks.

$$m_{ij} = x_{\left[n_{j-1}+0.5\left(n_j - n_{j-1}\right)\right],i}; \quad i = 1,2,\ldots,m; \quad j = 1,2,\ldots,k \tag{40}$$

*Option 3: Gaussian distributions.*

For $n \to \infty$ with one Gaussian variable the cut points for the optimal partition of a data set into $k=2,3,\ldots,6$ clusters have been computed by Cox (1967) under the condition that

$$\sum_{i=1}^{k} p_i \left( \frac{\mu_i - \mu_j}{\sigma_j} \right)^2 = maximum \tag{41}$$

where $\mu_j$ and $\sigma_j$ are, respectively, the mean and the standard deviation of the j-th variable, $\mu_i$ is the i-th conditional mean of $X_j$ given $L_i \le X_j \le U_i$, $i=1,2,\ldots,k$ and $p_i$ denotes the probability of an observation falling in the i-th group. I have extended the work of Cox for $2 \le k \le 25$.

| k | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | (-∞,0) | (0,∞) | | | | | | | | | |
| 3 | (-∞,-0.61) | (-0.61,0.61) | (0.61,∞) | | | | | | | | |
| 4 | (-∞,-0.98) | (-0.98,0) | (0,0.98) | (0.98,∞) | | | | | | | |
| 5 | (-∞,-1.24) | (-1.24,-0.38) | (-0.38,0.38) | (0.38,1.24) | (1.24,∞) | | | | | | |
| 6 | (-∞,-1.45) | (-1.45,-0.66) | (-0.66,0) | (0,0.66) | (0.66,1.45) | (1.45,∞) | | | | | |
| 7 | (-∞,-1.61) | (-1.61,-0.87) | (-0.87,-0.28) | (-0.28,0.28) | (0.28,0.87) | (0.87,1.61) | (1.61,∞) | | | | |
| 8 | (-∞,-1.75) | (-1.75,-1.05) | (-1.05,-0.50) | (-0.50,0) | (0,0.50) | (0.50,1.05) | (1.05,1.75) | (1.75,∞) | | | |
| 9 | ±0.22 | ±0.68 | ±1.20 | ±1.87 | | | | | | | |
| 10 | 0.00 | ±0.40 | ±0.83 | ±1.32 | ±1.96 | | | | | | |
| 11 | ±0.01 | ±0.41 | ±0.84 | ±1.33 | ±1.97 | | | | | | |
| 12 | 0.00 | ±0.01 | ±0.41 | ±0.84 | ±1.33 | ±1.97 | | | | | |
| 13 | ±0.01 | ±0.02 | ±0.42 | ±0.85 | ±1.34 | ±1.98 | | | | | |
| 14 | 0.00 | ±0.01 | ±0.02 | ±0.42 | ±0.85 | ±1.34 | ±1.98 | | | | |
| 15 | ±0.01 | ±0.02 | ±0.03 | ±0.43 | ±0.85 | ±1.34 | ±1.98 | | | | |
| 16 | 0.00 | ±0.01 | ±0.02 | ±0.03 | ±0.43 | ±0.85 | ±1.34 | ±1.98 | | | |
| 17 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.44 | ±0.86 | ±1.35 | ±1.99 | | | |
| 18 | 0.00 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.44 | ±0.86 | ±1.35 | ±1.99 | | |
| 19 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.45 | ±0.87 | ±1.35 | ±1.99 | | |
| 20 | 0.00 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.45 | ±0.87 | ±1.35 | ±1.99 | |
| 21 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.06 | ±0.46 | ±0.88 | ±1.36 | ±2.00 | |
| 22 | 0.00 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.06 | ±0.46 | ±0.88 | ±1.36 | ±2.00 |
| 23 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.06 | ±0.07 | ±0.46 | ±0.88 | ±1.36 | ±2.00 |
| 24 | 0.00 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.06 | ±0.07 | ±0.46 | ±0.88 | ±1.36 | ±2.00 |
| 25 | ±0.01 | ±0.02 | ±0.03 | ±0.04 | ±0.05 | ±0.06 | ±0.07 | ±0.08 | ±0.46 | ±0.88 | ±1.36 | ±2.00 |

Table 1: optimal grouping for a gaussian random variable

The values in table 1 are accurate to two significant digits. By using these cut points the partial means of the *m* variables are computed and inserted in a (kxm) matrix of typical values. If the first or the last interval were left empty then their mean is set equal, respectively, to the maximi and to the minimum of the variables.

*Option 4: "natural classes"*

Mineo (1985). Let $\{x_{j,r}, n_{j,r}, r=1,2,\ldots,m_j\}$ be the frequency distribution of the j-th variable where the values are sorted by size and $m_j$ denotes the number of distinct values observed for $X_j$. Determine the minimum of

$$D_r = \left(\frac{n_{j,r}n_{j,r+1}}{n_{j,r}+n_{j,r+1}}\right)\left(x_{j,r+1}-x_{j,r}\right)^2, \quad r=1,2,\ldots,m_j-1 \tag{42}$$

The value of $x_{j,r}$ is replaced by

$$x^*_{r,j} = \frac{x_{j,r}n_{j,r}+x_{j,r+1}n_{j,r+1}}{n_{j,r}+n_{j,r+1}} \qquad with \quad n^*_{r,j} = n_{j,r}+n_{j,r+1} \tag{43}$$

The value $x_{j,r+1}$ is eliminated and the values above are shifted back to form a new frequency distribution with $m_j$ *-1* values. These two steps are iterated until the frequency distribution has only *k* distinct values. The same procedure is repeated for the *m* variables to define the matrix of *(kxm)* representative values. A serious drawback of all these methods is that, as dimensionality increases, the volume of the data concentrates in the external boundary with the consequence that high dimensional space is mostly empty which, in turn, implies that these methods are doomed to find most of the partitions invalid because one or more clusters have no entities in it. To avoid invalid partitions, random entities are selected from the largest clusters and placed in the empty clusters.

### 3.2.3 Random combinations of entities.

Let $P=\{P_1, P_2, ..., P_k\}$ a random samples without replacement of *k* entities from the data set of *n* entities; then each entity is assigned to its closest centroid according to (22) (this ensures that each cluster contains at least one entity). The procedure is repeated until $Min\{\nu, C(n,k)\}$ partitions are examined. In particular, if $\nu > C(n,k)$ then all possible combinations of *n* entities taken *k* at a time are considered as initial centroids. However there is no guarantee that a "true" centroid coincides with one of the entities to cluster so that even a complete enumeration of all combinations may result in an inappropriate initial partition. This method presents a similar problem to that examined in section 3.2.1. In fact, when two or more of the selected entities are close together so that there will be two or more cluster close together which not necessarily are present in the data set. In addition, if clusters are of unequal size, the small cluster have lower chances to generated a centroid and tend do be absorbed by the larger ones.

To remedy this shortcoming ***DetClus*** considers $Min\{4k, n/2\}$ randomly selected and distinct entities and choice the best *k* centroids by applying the Kennard_Stone procedure (*option 5).*

### 3.2.4 Random partitions

A set of *n* integers is chosen as follows

$$\gamma_r^0 = j \quad if \quad P_{j-1} \le \varpi_r \le P_j, \quad j = 1,2,\ldots,k; \qquad r = 1,2,\ldots,n \tag{44}$$

where $\omega_r$ is a pseudorandom numbers from *[0,1]*. The quantities $P_0, P_1, ..., P_k$ are given by

$$P_0 = 0, \quad P_t = \sum_{i=0}^{t} v_{ri}; \qquad v_{ri} = \frac{v^*_{ri}}{\sum_{j=1}^{k} v^*_{rj}}, \quad i = 1,2,\ldots,k \tag{45}$$

where $v^*_{ri}$ *i=1,2,...k* are random numbers from *[0,1]*. The previous expressions ensure that each cluster always contains at least one entity.

### 3.2.5 Random shuffling

Let $\gamma=(\gamma_1, \gamma_2, \ldots, \gamma_n)$ be a vector of integers between 1 and *k*. By using the technique suggested by Knuth (1981, p. 139) random permutations of the $\gamma_r$ 's are considered. The set of numbers to be shuffled is chosen as follows

$$\gamma_r^0 = j \quad for \quad r = N_{j-1}, N_{j-1}+1, \ldots, N_j - 1; \quad where \quad N_j = \sum_{i=0}^{j} n_j; \; n_0 = 1, \; j = 1,2,\ldots,k$$

The user must specify the cardinalities of the clusters. This option allows the algorithm to explore the partitions with the same number of members per cluster.

### 3.3 Applications of the Indifference Principle

Since we ignore the real cluster membership of the entities, each entity should have the same chances of joining one of the *k* cluster. An initial configuration based on this policy is free of overt biases. Le *b=[n/k]* and *s=n-b*k*;

*Option_1: equal membership partition*

Each cluster has the same number of entities except the last group which is assigned all extra entities. To obtain such a partition the first *b* entities are assigned to cluster $C_1$; entities labelled from *b+1* to *2b* to cluster $C_2$ and so on. The last *s* entities are added to the last cluster.

*Option_2: discrete uniform distribution*

For each entity *r* a random number *j* is generated from the discrete uniform *[1,k]* distribution and $\gamma_r^{(0)}=j$ for *r=1,2,...,n*. The empty clusters receive an entity from a regular cluster

*Option_3: random blocks*

Step_1. Set $\gamma_r=0$ for $r=1,2,\ldots,n$. Set $h=1$.

Step_2. Generate $b$ distinct random integers $u_i$, $i=1,2,\ldots,b$ in the interval $[1,n]$.

Step_3. Set $r=u_i$. If $\gamma_r=0$ then assign entity $X_r$ to cluster $C_h$. Set $\gamma_r=1$.

Step_4. If $h<k-1$ then set $h=h+1$ and go to Step_2.

Step_5. If $\gamma_r=0$ then assign entity $X_r$ to cluster $C_k$ for $r=1,2,\ldots,n$.

*Option_4: nested loops*

The entities labelled $\{j, k+j, 2k+j, \ldots, (b-1)k+j\}$ are assigned to the cluster $C_j=j=1,2,\ldots,k$. The last $s$ entities are added, one for each, to the first $s$ clusters.

### 3.4 Read centroids from file

The user can provide the estimated centroids from a text file in which the rows are the centroids and the columns are the variables. This options is allowed for a fixed number of clusters. The program checks the internal conditions: $L_j \le \mu_{ij} \le U_j$, $j=1,2,\ldots,m$; $i=1,2,\ldots,k$. If this condition is not satisfied then each invalid entry is replaced by a uniform random number in the interval $[L_j,U_j]$. The corresponding classification vector is obtained by applying (22). The partition is discarded if some cluster is empty.

### 3.5 Read partition from file

Sometimes the entities to be clustered have *a-priori* labels and one is investigating wether the cluster membership that can be obtained by the algorithm is consistent with the known labels (supposing these to be a plausible classification of the data set). Moreover, this option allows the user to start **DetClus** from the configuration achieved by another procedure (internal or external to the program).

The program accepts the proposed partition only if each clusters has at least one empty. The number of clusters is derived from the number of distinct labels found in the file.

$$\left[\gamma_1 \quad \gamma_2 \quad \cdots \quad \gamma_i \quad \cdots \quad \gamma_n\right]$$

## 4.The quality of a partition

Any clustering algorithm constructs a partition $\gamma^* \in P(n,k)$ which is optimal in terms of the stated criterion and the initial solution, with as many clusters as desired (virtually every definition of optimal clustering does not depend on the number of clusters). However, the clustering found will be useful only if the classes can be substantively interpreted. Fisher and Van Ness (1971) observed that the main objective of a clustering is to condense information by reducing the individual description of all X's to a relatively few general description of $k$ typical representatives, one for each cluster. Paradoxically, if the variables were constant within the clusters, one entity per cluster would suffice to express any detail of the data set. As a rule, the lower $k$ is the stronger is the partition since less information is needed to summarize the data; hence, when there is more than one optimal solution, the one with the lower number of clusters should be chosen. Castagnoli (1977) has shown that such a partition always exists. The problem is further compounded by the fact that, as we have seen in the previous section, the number and the type of clusters in the data may depend on the resolution with which we look at the data.

One major problem shared by all methods of cluster analysis is that an optimal partition of the data set into a certain number of nonempty subsets with pairwise empty intersections will be developed whether or not a natural clustering exists and whether or not it is possible to select plausible centroids among the data set.

***Example:***

In dissection, the data set comprises entities whose distribution into the space of variables is uniform; the aim is to subdivide the entities into sectors (e.g. Policy precincts, voting districts, school districts and so forth). Nevertheless, it is legitimate to wonder whether entities in different sectors of Figure 11 are heterogeneous and whether the clusters obtained have a real existence.

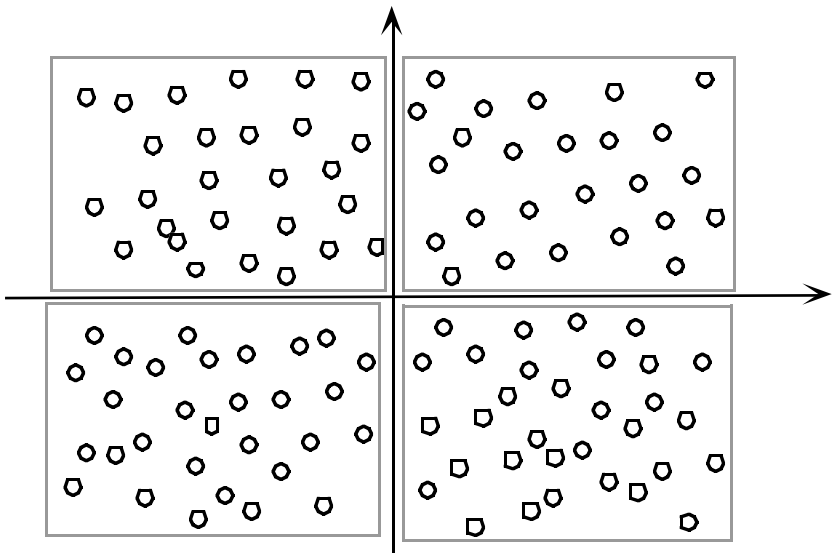


Figure_11: artificial clustering

The quality of a clustering is partly intrinsic to the data-generating process, the data collection equipment, the choice of the variables, and a possible selective identification of the entities to be clustered. These issues are, however, outside the scope of the present section; here the intent is to devise extrinsic aids (graphical and numerical) for distinguishing meaningful partitions from those artificially imposed on the entities.

***Example:***

This experiment was constructed by simulating points from 3-dimensional random variables having uniform marginal distributions. Let $\boldsymbol{u}_i$ be a vector of $m$ independent uniform random variables on $(0\text{-}1)$ with $E(\boldsymbol{u}_i)=0.5\,\boldsymbol{c}_m$ and $E(\boldsymbol{u}_i\boldsymbol{u}_i^t)=(12)^{-1}\boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix of order $m$; then $\boldsymbol{Y}_i=\sqrt{12}(\boldsymbol{u}_i-0.5\boldsymbol{c}_m)$ is a vector of independent uniform random variables with $E(\boldsymbol{Y}_i)=\boldsymbol{0}$ and $E(\boldsymbol{Y}_i\boldsymbol{Y}_i^t)=\boldsymbol{I}$. Consider now the affine transformation $\boldsymbol{X}_i=\boldsymbol{H}\boldsymbol{Y}_i+\boldsymbol{d}_i$ where $\boldsymbol{H}\boldsymbol{H}^t$ is the Cholesky factorization of $\sum$; then $\boldsymbol{X}_i$ *is a a* m-dimensional random variables having uniform marginal distributions with $E(\boldsymbol{X}_i)=\boldsymbol{d}_i$ and variance-covariance matrix $E(\boldsymbol{X}_i\boldsymbol{X}_i^t)=\boldsymbol{H}\boldsymbol{H}^t=\sum$ Usually, only synthetic data sets include "natural" clusters exhibiting high level of external isolation and internal cohesion. However, if one encounters such data (and there is no reason to suspect an happenstance, an error or a joke), it would not be hard to find a convincing *post hoc* rational explanation which legitimates the empirical results.

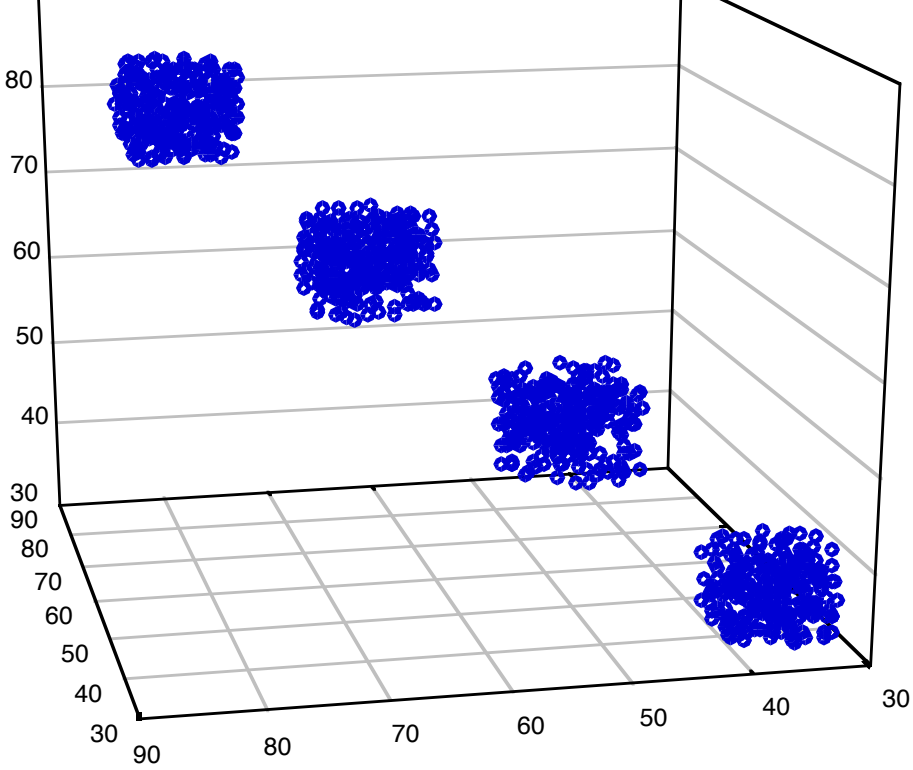


Figure 12: ideal data set

Though the partitioning of natural clusters appears meaningful and potentially useful, the partitioning of a unimodal or uniformly distributed entities does not appear to have the same basis (Arnold, 1979).

***Example***

Späth data set (Späth, 1985, p.144).
Two variables for 41 entities randomly scattered over the variable space without accumulation zones; none of the interpoint distances is significant; there is no natural grouping within the data so that any rule proposing a "plausible" partition into $k$ groups should be critically exhamined. ***DetClus*** for $k=3$, has produced an arbitrary dissection (figure13) along the axis of maximal dispersion. Marriott (1971) noted that minimization of $Min\{|\boldsymbol{W}(\gamma)|\}$"... *searches for any natural grouping, not necessarily one based on all the measurement*".

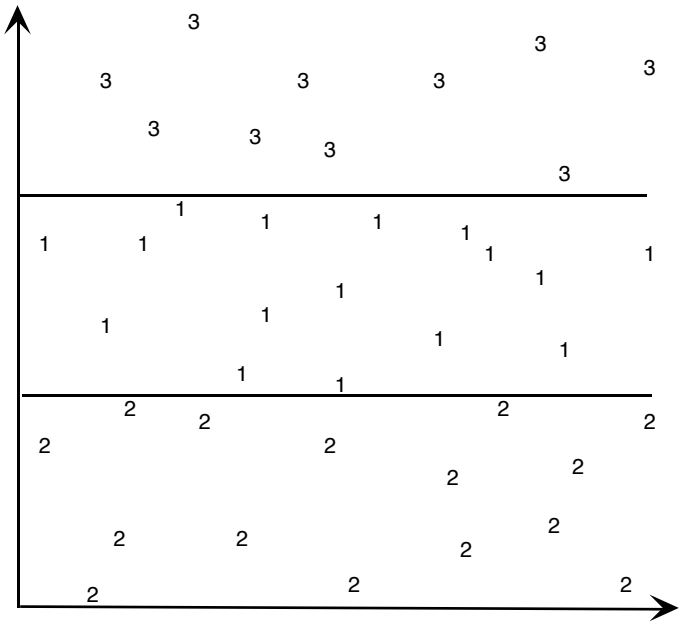


Figure 13: results for the Späth data set

***Example***

Unimodal data set. A sample of 250 bidimensional entities uniformly distributed within the ellispoidal region $\boldsymbol{x'\Omega^{-1}x \le 1}$ where

$$\Omega = \begin{bmatrix} 9 & 3 \\ 3 & 9 \end{bmatrix}$$

***DetClus*** has no protection against finding groups in the data when in effect none is present. For $k=3$, it finds a seemingly plausible partition which is actually unexplicable in terms of what is known on the data. The fact that the clustering algorithm has found a structure doed not imply that ther sdtructure is real.

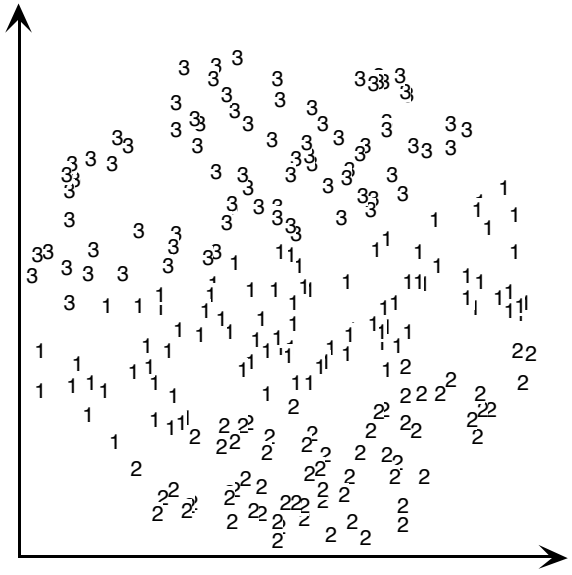


Figure 14: results for the unimodal data set

In both examples it appears difficult to argue that one particular solution has more meaning or stability in either a logicakl or theoreticalk sense that any other clustering that can be randomly generated.

In general, each clustering is a good clustering if there is theoretical and circumstantial evidence that may convincingly explain the structure obtained; conversely, any clustering, optimal though it may be, lacking an explanation as to how the member of a group came to be described as similar, and how these members differ from those of other groups, is merely an artifact of the algorithm.

Any expert or practitioner of cluster analysis knows that the output of a clustering procedure is not the end of the story, but several questions must be answered. Bock (1995) suggests the following

1) What is the relevance and significance of the resulting classes?
2) Do they reflect a "true" or "natural" grouping structure of the data or just an artifact of the method selected?
3) How does the clustering perform when compared to random classifications?
4) Which are the strongest or the most doubtful classes?

In general, procedures used to evaluate clusters determined by a clustering method are of two types. The first one includes procedures for testing the resultant clusters against the null hypothesis that the clusters were randomly determined. Procedures of the second type are based on the assumption that the clustering method in use has attained an optimal partition which is compared with a given partition for comparison purposes.

### 4.1 External indices of validation

The effectiveness of relocation procedures can be measured by comparing the final partition $\gamma^+$ generated by the algorithm with the prior knowledge of the true classification $\delta$. Sometimes the iterative scheme is starting from $\delta$ which should be, hopefully, in the domain of attraction of a global minimum. This situation is very unrealistic in that it tacitly assumes that not only the number of clusters, but also the true cluster membership of all $n$ entities is known. However, such idealized setting offers a simple benchmark against which the results can easily be compared. In particular ***DetClus*** computes the Hubert-Arabie (1985) statistic

$$R_{HA} = \frac{nc(n-1)(c-1) - ab}{n(n-1)b - ab};$$

$$a = \sum_{i=1}^{k} n_i^+ \left( n_i^+ - 1 \right);\ b = \sum_{i=1}^{k} n_i \left( n_i - 1 \right);\ c = \sum_{r=1}^{n-1} \sum_{s=r+1}^{n} \varepsilon\left( \gamma_r^+ = \gamma_s^+ \cap \delta_r = \delta_s \right) \tag{46}$$

where $\varepsilon(x)$ is one if $x$ is true and zero otherwise. The statistic $R_{HA}$ has a fixed upper bound $R_{HA} = 1$ indicating perfect clustering recovery and takes the value zero under the hypothesis that $\gamma^+$ and $\delta$ are picked at random subject to having the true number of clusters and objects in each. In

addition, ***DetClus*** computes a naive index of clustering efficacy

$$Q = \frac{\pi_C - \pi_L}{1 - \pi_L}; \qquad \pi_C = \sum_{i=1}^{k} \binom{n_i(\gamma^+)}{2} \Big/ \binom{n}{2} \; ; \qquad \pi_L = \sum_{i=1}^{k} \binom{n_L(\delta)}{2} \Big/ \binom{n}{2} \tag{47}$$

where percentage $\pi_C$ is the proportion of pairs in which the two entities are in the same clusters both in $\gamma^+$ and $\delta$ while $\pi_L$ is the percentage of pairs of entities belonging to the largest cluster of δ. In pracgtice, the statistic$Q$ compares the goodness of the classification resulting from a k-means algorithm and the naive classification obtained putting all the entities in one cluster. A negative value of $Q$ indicates that ***DetClus*** was not able to detect any clustering in the data set, at least for the given starting partition. The user must be aware that (46) and (47), as well as, many other external indices of agreement, are not a naturally increasing function of the quality of the partition found by the procedure.

***Example:***

Ruspini data set. (Kaufman and Rousseeuw, 1990, p. 100). This is a standard example consisting of 75 two-dimensional points making up *k=4* natural groups including 23, 20, 17, 15 entities . Actually the these data are different form the original data used by Ruspini (Rasson e Kukbushi-shi,1994, p. 191).

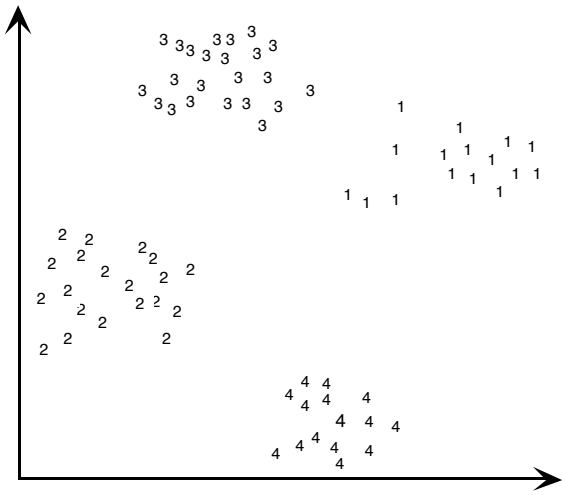


Figure 15: results for the Ruspini data

In this example the groups are well-structured and any reasonable method of cluster analysis can isolate them. ***DetClus*** does not fail to retriew this obvious structure: $R_{HA}=1$ and $Q=1$.

***Example:***

Storm survival of sparrows (Bumpus data set). After a severe storm on 1 February 1898, a total of 136 sparrows *(Passer domesticus)* were taken to Bumpus's laoratory. Bumpus took m=9 morphological measurement on each bird and also weighted them. Manly (1985) reproduced his data classified according to sex and the age of males for a total of six clusters having the cardinalities: young males that survided=16; young males that died=12; adult males tha survided=35; adult males that died=24; adult and young females thad survided=21; adult and young females that died=28.

The correlation matrix is positive (each elements is greater than zero) so that the first principal component is an index of size (factor loadings having the same sign and roughly equal magnitude) whereas the other components are contrast or shape components (at least one factor loading has a sign different form the others).

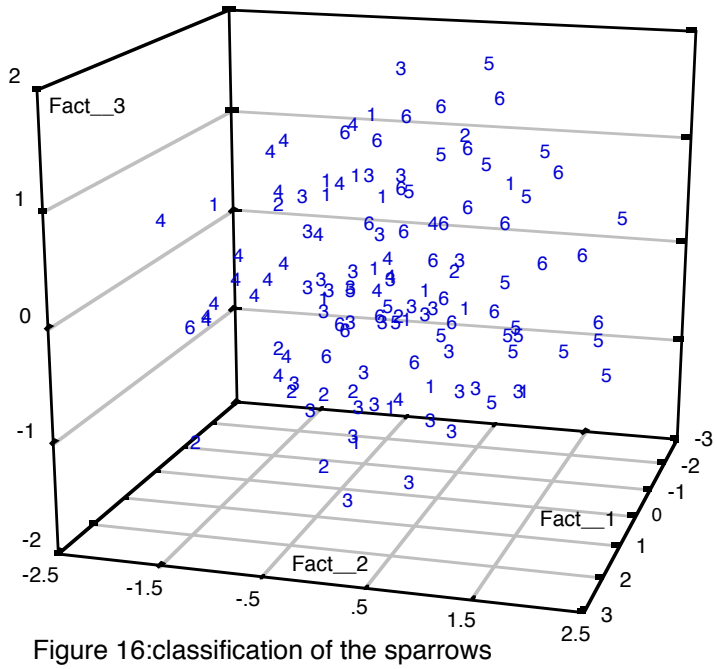



Figure 16:classification of the sparrows

The space of the first three PC's, which explains the 76.6% of total variation , does not show any particular structure. For $k=6$ **DetClus** found $R_{HA}=0.075$ and $Q=-0.012$ The quality of the results does not improve when the final partition of **DetClus** is compared with the subdvision of the data set according to the sex or according to survivors/non survivors sparrows.

### 4.2 Estimation of the number of clusters

There is no standard way of statistically evaluating the adequacy of the obtained sequence of partitions. The vagueness of the theoretical basis makes it difficult to achieve analytical results in this area and preference should be given to graphic displays. These techniques are very useful in the validation of a clustering even though it has proven unreliable to trust intuition or visual perception alone. Blanshfield *et al* (1982) have observed that iterative partitioning algorithms are much better than hierarchical algorithms concerning output descriptive statistics to making it possible to obtain many graphic views for more intimately inspecting the clustering process.

#### 4.2.1 Complete clustering characteristic graph

One of the most popular methods of choosing the appropriate number of clusters is to plot the objective function against the number of clusters $k$ for a range of values of $k$. The true number of clusters is found by considering those values of k fro which the plot shows a sharp in/decrease of the criterion. ***DetClus*** considers two indicators

$$Friedman-Rubin: \quad C = \frac{Min\left\{\left|\boldsymbol{W}(i)\right|\right\}}{|\boldsymbol{T}|} * 100; \qquad i = k_1,\ldots,k_2 \tag{48}$$

The criterion is normalized by the corresponding value for $i=1$ so that (48) lies in the interval $(0,100)$. Expression C is a decreasing function of the number of clusters and an increasing function of the number of entities and dimensions. Undoubtedly, with every increase in $i$ there will be a decrease in (48), but the change should be irrelevant for $i>k$ when $k$ is the number of cluster which best fits the data. In practice, a discontinuity in slope should correspond to the true number of clusters, otherwise there no justification for having more than one class (Hardy,1996).

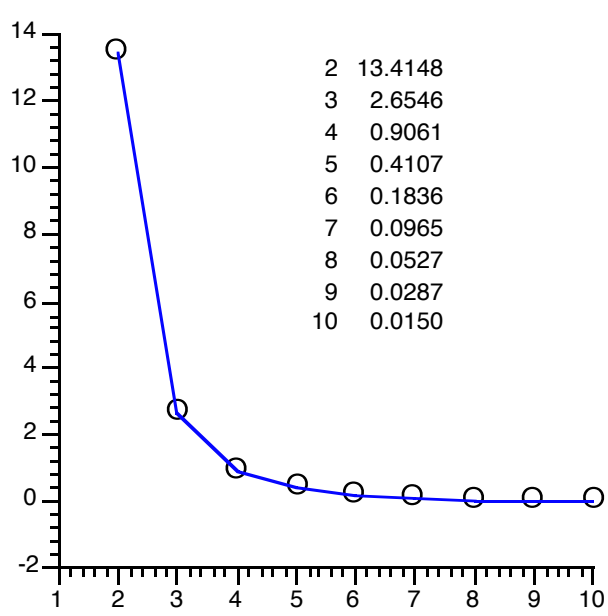
Arnold (1979) proposed the following test statistic

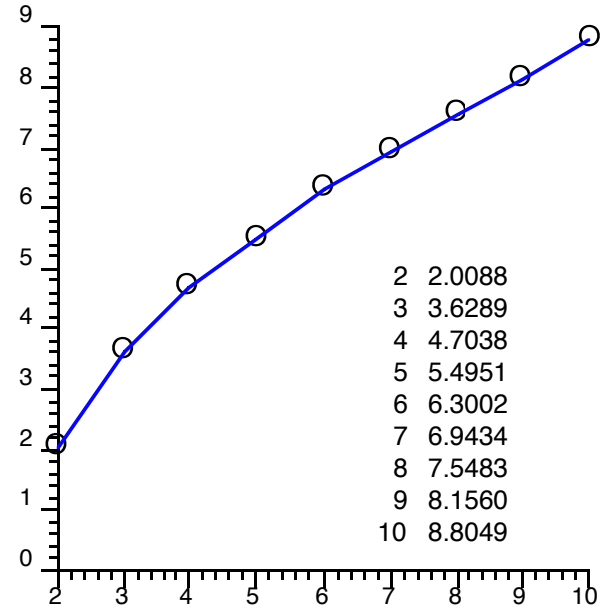
$$\alpha = Ln\left\{\frac{|\boldsymbol{T}|}{|\boldsymbol{W}(\boldsymbol{i})|}\right\}; \qquad i = k_1,\ldots,k_2 \tag{49}$$

for testing the null hypothesis that the entities are either uniformly distributed or grouped into clusters. The method of deriving the distribution of $\alpha$ was based on Monte Carlo techniques, but the results are not satisfactory. However, the plot of (49) can be used as (48) to correctly estimate the number of clusters. ***DetClus*** writes (48) in the output file, but the user can easily compute (49) by $\alpha=Ln(100/C)$. The user must be aware that highly collinear variables can create problem to the Anderson statistic.

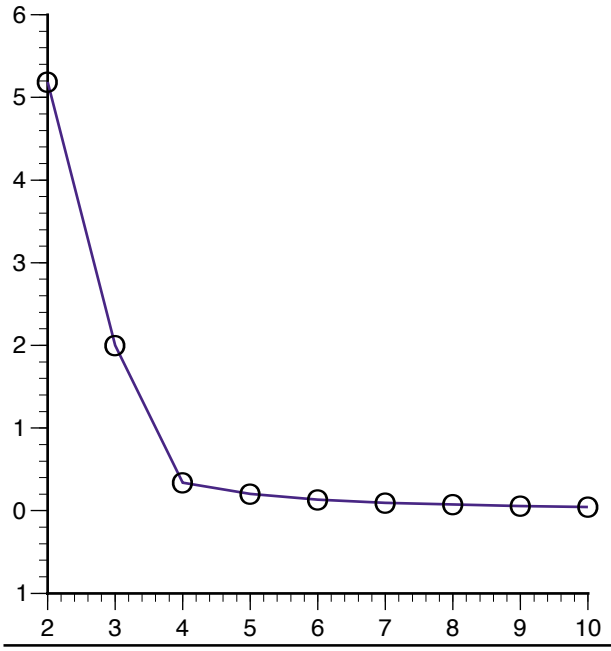
***Example***

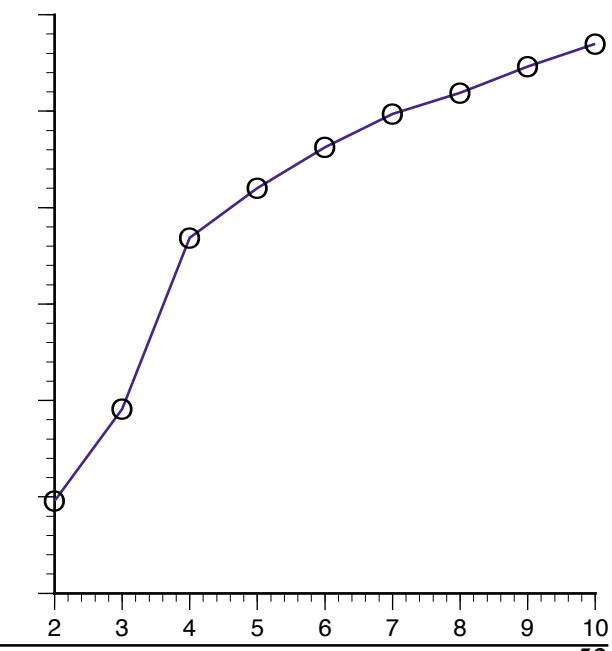
The statistics of poverty and inequality (Rouncefield, 1995). For $n=97$ countries in the world, data are given for birth rates, death rates, infant death rates, life expectancies for males and females, and GNP. For this example the first four principal components were used (98.9% of total variation explained). The clustering of the data set appear to be weak and does not correspond to the classification in $k=6$ clusters proposed by the geographical grouping included in the data. The value $k=4$ is a plausible choice because of the sharp decrease noted in (48) and the progressively reduced increments in (49) after $i=3$, but other choices can easily be made.




For the Ruspini data sets, both the indices perform well.



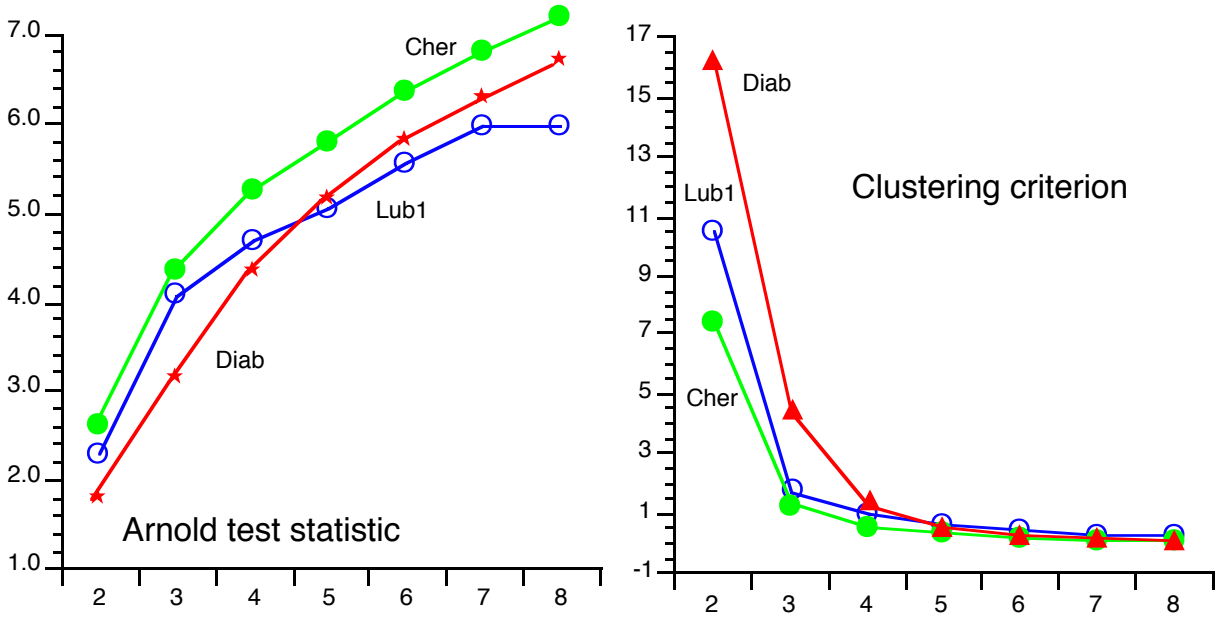


***Examples:***

1) Lubishew data set 1 (Lubishew, 1962). Measurements were made of six variables in the males of three species *Chaetocnema concinna, Ch. heikertingeri, and Ch. heptapotamica,* The real composition of the groups is (21, 31, 22). ***DetClus*** correctly assigned to the appropriate cluster all the entities even though only the first three principal components (89.3% of total variation) were used to identify the specimen.

2) Fossils data (Chernoff, 1973). Six variables were measured on each of nummulited specimens from Eocene Yellow Limestone formation of Northwestern Jamaica. According to Chernoff the entities divide into three distinct clusters: {40, 34, 13} with one or two specimen which can be regarded as singleton or borderline. ***DetClus*** has been applied to the first four principal components (accounting for 94.6% of the variability contained in the data set) providing perfect recovery of all the entities. However, the largest cluster can be separated into subclusters, but their number is undeterminate.

3) Chemical and overt Diabetes (Andrews and Herzberg,1985). This data set consists of five variables (insulin area, glucose area, and steady-state plasma glucose response) measured on *n=145* non obese adult subjects. The subjects were clinically classified as normal (76), Chemical diabetes (36) and overt diabetes (33). The clusters have various sizes and different non-ellipsoidal dispersion matrices.




For the Lubishew1 data set *k=3* is an evident point of inflection. The Chernoff data set shows a drop (or a jump if you are looking to the Arnold statistic) at *k=3* and at *k=4* but it is non easy to make a decision without further analysis. The graph for the diabetes data set is confused. However, a partition in *k=3* or *k=4* cluster is deemed to be plausible.

***DetClus*** provides a very raw graph of (48), but the value of the criterion can be copied from the output file and pasted in one's favorite plotting program (Excel, Statistic, Deltagraph, etc.£

Hall and Khanna (1977) have great confidence on this type of graph: a knee (i.e. a sharp step from $i$ to $(i+1)$ followed by a marked flattening of the curve suggests that $k=i+1$ is a good choice. Other authors (*e.g.* Everitt, 1979; Gordon, 1999, p. 61) do not recommend great reliance on this graph. Three good reasons for such reservations are:

a) A data set often exhibits more than one point of diminishing return (that is, the value of $k$ at which the rate of decrease in the slope starts to diminish) and it is difficult to tell which indicates the correct number of clusters.

b) Frequently the plot has a knee even if the conjoint cluster solution might be considered the best partition.

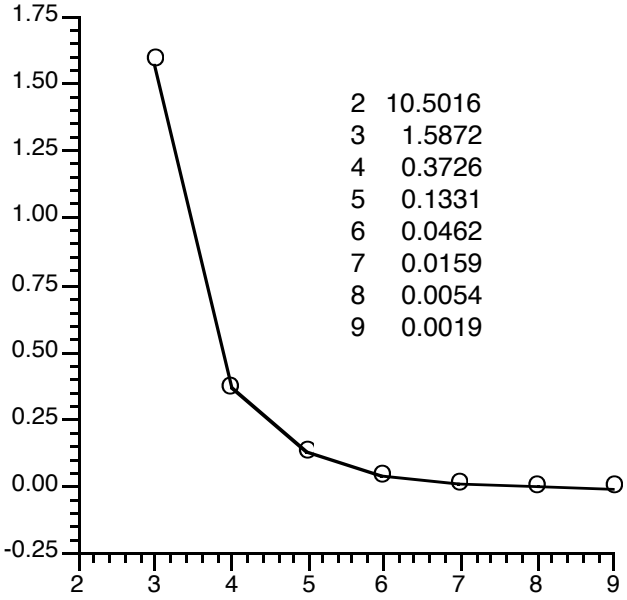
c) It may be difficult to locate the critical point in the graph for large values of $k$ where the variations are small anyway.

The above-mentioned problems are frequent when the structure of the data set is very complicated. Unfortunately, these are just the occasions when an effective means for comparing alternative clusterings becomes more acutely necessary.

The plots by themselves do not rigorously reveal how many clusters are actually present. Rather, they are useful guidelines in selecting an appropriate number of clusters in a context where developing inferential methods has proved difficult. The defects of a subjective estimation of $k$ result, in the main, from uncertainty which is demonstrated when different observers have to decide on the same plot and obtain different answers. In fact, Milligan and Cooper (1985) excluded from their review any technique requiring human judgement, but took into consideration analogous procedures, based on "difference scores", that are not so different from visual assessments.

*Example:*

Egyptian skulls data set (Hand *et al.* 1994). Four measurements of male Egyptian skulls from five different time periods. Thirty skulls are measured from each time period ($n=150$). The elimination of the point corresponding to $k=2$ evidentiates the knee at $k=5$ (the true value of the number of clusters). However, the recovery rate is extremely poor: $R_{HA}=0.0041$ and $Q=-0.09$.



### 4.2.2 Stopping rules

The validation stage of cluster analysis has a relevance of its own and specific test for clustering quality should be explored. In particular, guidance dependent only upon the actual partition of the entities and not on the algorithm used to obtain them should be provided by any efficient clustering program. A weakness of the complete characteristic curve is its tendency to mask comparisons between partitions since it summarizes a clustering by a single number. No single plot is likely to convey all the relevant features of a partition. Many other indices can be computed and plotted, but they must be sensitive, informative, easily computed with a single scan of the entities and consistent with the algorithm used in obtaining the clustering (for instance, statistics based on the Euclidean metric are inapplicable to ***DetClus***).

Any conceivable function for measuring clustering adequacy should be able to detect the following situations:

a) Conjoint cluster.

This definition refers to a strongly unimodal distribution of the entities which are concentrated around a single center that is, all the distances between entities and the total mean are below a critical value. The common denominator of these two situations is the fact that no configuration can be considered better than any other. The only plausible representation of the data is a partition in which all the entities are in the same cluster.

b) Exactly *k* cluster.

The variable space shows *k* high density regions completely surrounded by gaps. Any attempt to classify the entities in a number of clusters lower than or greater than *k* would create fictitious clusters.

c) *n* clusters (disjoint partition)

Entities are randomly scattered over the variable space without accumulation zones; each entity forms a single point cluster .

As we have seen, ***DetClus*** cannot detect situation a) and situation c). However, the recovery rate of ***DetClus*** is fairly satisfactory and it may be expected that the program is capable to reconstruct the true underlying cluster structure (if a cluster structure is actually present in the data set).

Reviews of procedures for determining the number of clusters are Sadocchi (1977), Bailey and Dubes (1982), Dubes (1987), Vicari (1990), Hardy (1996), Gordon (1996). In particular Milligan and Cooper (1985) presented a fairly extensive coverage of the so-called stopping rules: thirty procedures were evaluated by Monte Carlo simulations, and, although findings are likely to be somewhat data depending, the general quality of results is encouraging. A serious reservation about the study is the fact that stopping rules for hierarchical and non hierarchical

methods were jointly treated. As it is well known (perhaps is the only non controversial point in cluster analysis) the two approaches pursue different purpose. Hierarchical methods produce a series of solutions ranging from $n$ clusters of size one to one cluster of size $n$ *(or vice versa)*; for iterative partitioning methods the number of clusters is a parameter fixed in advance. Moreover, in hierarchical schemes, an entity is indissolubly tied with entities in the same group; in iterative methods, each entity at each stage, is free of moving from one cluster to any other.

Dubes (1987) noted a close link existing between the method of clustering and the performances of a stopping rule. The two essential questions with re-allocative processes are:
a) The lack of guarantee that algorithms will arrive at the absolute optimum partition. This puts additional difficulties on the estimation of the number of clusters since another source of error has to be taken into account.
b) The number of variations with which a method can be carried out, and, as a result, the absence of a generally accepted formalism for comparisons.

These issues does not seem to have been fully exploited by Milligan and Cooper (1985) thus missing the fact that a method of cluster analysis is concerned with determination of the optimal number of classes as well the optimal classification for that numbers. If the user is sure that his/her dataset contains clusters the most appropriate value of $k$ compatible with the clustering procedure can be assessed by considering many different statistics.

1) Pseudo F Statistic (Calinski, 1969)

$$K(i) = \left(\frac{n-k}{k-1}\right) \frac{Trace\left(\boldsymbol{T^{-1}B}\right)}{Trace\left(\boldsymbol{T^{-1}W}\right)} \tag{50}$$

A monotonic increasing sequence of the pseudo F-statistic suggests that there exist no well separated clusters, while a monotonic decreasing sequence can be expected in the presence of a hierarchical structure. The global maximum may be regarded an indication for the "best number of cluster", though we may prefer to divide the data set into the number of clusters suggested by a local maximum of *K(i)* (Calinsky, 1969).

The literature on cluster analysis usually considers the statistic

$$K(i) = \left(\frac{n-k}{k-1}\right) \frac{Trace(\boldsymbol{B})}{Trace(\boldsymbol{W})} \tag{51}$$

suggested by Calinski and Harabasz (1974). The two indices coincide if the variables are the principal components of the data set. If the two rules give the same indication then there is substantial evidence that the data set has the given number of cluster. If (50) and (51) do not

show a local minimum or a local maximum for the same value of k then either the initialization method is inadequate or no significant clustering is present in the data or, if present, cannot be recognized by **DetClus** .

2) The C-index.
An index reviewed by Huber and Levine (1976) and discussed by Gordon (1996) for assessing a partition into *k* clusters of a set of *n* entities is based on the sum of all within-cluster pairwise distances

$$C(i)=\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}\delta_{rs}M(X_r,X_s),\quad \delta_{rs}=\begin{cases}1 & if\ \ \gamma_r=\gamma_s=j;\\ 0 & otherwise\end{cases}\quad M(X_r,X_s)=(\boldsymbol{X}_r-\boldsymbol{X}_s)^t\boldsymbol{W}^{-1}(\boldsymbol{X}_r-\boldsymbol{X}_s) \quad (52)$$

If the number of entities *n* is small then $D_{min}$ (resp., $D_{max}$) is defined as the sum of the $N$ smallest (resp., largest) pairwise distances and (51) is standardized as

$$C^*(i)=\frac{C(i)-D_{min}}{D_{max}-D_{min}};\qquad where\ \ N=\sum_{j=1}^{k}\binom{n_j}{2} \quad (53)$$

The value $i=k$ which minimizes $C^*(i)$ is regarded as specifying the number of clusters in the data set. For large data set $C(i)$ can be standardized by

$$C^*(i)=\frac{\frac{C(i)}{N}-E_{min}}{E_{max}-E_{min}};\quad E_{min}=\underset{1\le r,s\le n,r\ne s}{Min}M(X_r,X_s),\ E_{max}=\underset{1\le r,s\le n,r\ne s}{Max}M(X_r,X_s) \quad (54)$$

although, in this case, the link between the values of the index and the best value of *k* deserves further study.

3) Marriott index

$$M(i)=i^2\frac{|\boldsymbol{W}|}{|\boldsymbol{T}|} \quad (55)$$

If the values of $M(i)$ lie in a restricted band there is no evidence of separated clusters. A real natural grouping would be shown by a point lying well below the general trend (Milligan and Cooper (1985) used the maximum difference between successive levels and ranked $M(i)$ 20-th best of thirty indices. If $M(i)\ge 1$ for all values of $i$, the data should be regarded as a conjoint cluster. The Marriott criterion has received serious criticism but it has been inserted into the present study because of its consistency with the objective function adopted by ***DetClus*** .

4) Silhouette coefficient (Struyf et al. 1997).

$$S(i)=\sum_{r=1}^{n} s_r;\quad s_r=\frac{b_r-a_r}{max\{a_r,b_r\}};\quad a_r=\frac{\sum_{\gamma_s=i} M(X_r,X_s)}{n_{i-1}}\ \ with\ \ \gamma_r=i;$$
$$b_r=\underset{j=1,..,k;j\neq i}{Max}\{\eta_{ij}\},\ \ \eta_{ij}=\frac{\sum_{\gamma_s=j} M(X_r,X_s)}{n_j}\ \ with\ \ \gamma_r=i\neq j \tag{56}$$

$a_r$ denotes the average distance from $X_r \in C_i$ to all other object in the cluster $C_i$; $b_r$ denotes the minimum among the average distances of $X_r \in C_i$ to all object in cluster $C_i$, $i=1,2,\ldots,k,;\ i\neq j$. When data set is well-structured the proper number of cluster determines a maximum of $S(i)$ near the upper bound $(S(i)\leq 1)$. If the silhouette coefficient is below 0.25 we may conclude that non substantial clustering structure is to be found in the data.

5) Dunn index

The validation of the clusters found is often based on the notions of internal cohesion within clusters or external cohesion of clusters. There are a variety of indices which attempt to take into account these two requirement which may be difficult to reconcile since one goal (internal cluster cohesion) can conflict with another (separation between clusters).

The Dunn index can be defined as

$$D_1(i)=\underset{e,g=1,2,..,i;e\neq g}{Min}\left\{\frac{\phi_{e,g}}{\psi_e}\right\};\ \phi_{e,g}=\sum_{\gamma_r=e}\sum_{\gamma_s=g} M(X_r,X_s),\ \psi_e=\sum_{\gamma_r=e}\sum_{\gamma_s=e} M(X_r,X_s),\ n_e>1 \tag{57}$$

where the sums are over $r,s=1,2,\ldots,n$. The Index (57) has been discussed by Bezdek and Pal (1998), Kotari and Pitts (1999). Usually a point of maximum is indicative of a good value for the number of clusters. Another version (Theodoridis and Koutroumbas, 1998,p.562) of the Dunn index is

$$D_1^*(i)=\underset{e,g=1,2,..,i;e<g}{Min}\left\{\frac{M(X_e,X_g)}{\Delta}\right\};\ \Delta=\underset{j=1,2,\ldots,k}{Max}\left\{\underset{\gamma_r=j,\gamma_s=j}{Max}\{d(\boldsymbol{X}_r,\boldsymbol{X}_s);\ r,s=1,\ldots,n\}\right\} \tag{58}$$

It is clear that if the clusters are compact and well-separated the index (58) will be large since the distance between the clusters is expected to be large whereas the diameter of the clusters tends to be small. A maximum in the plot of $D_1^*(i)$ can be used to indicate the number of clusters in the data set. Dunn suggested that if $D_1^*(i)>1$ then the clustering if formed by compact and well-separated clusters.

6) Cluster assessment statistics.

Klastorin, 1983 proposed a class of measures for testing wether resultant clusters differ significantly from randomly determined clusters of the same size: $g_r = w_r - b_r$ where $w_r$ is a measure of within-cluster homogeneity and $b_r$ is a measure of between-cluster heterogeneity. Since clustering methods attempt to minimize the within-cluster variance, methods that purport to test the final partition against the null hypothesis that entities are assigned randomly to clusters are useless. However, they provide a framework which can lead to an effective choice of *k*.

For example, the first measure proposed by Klastorin is

$$g_1(i) = \left(\frac{2}{i}\right)\left\{\sum_{e=1}^{i}\left[\frac{\psi_e}{n_e(n_e-1)}\right] - \sum_{e=1}^{i-1}\sum_{g=e+1}^{i}\frac{\phi_{e,g}}{(i-1)n_e n_g}\right\} \tag{59}$$

The first term is the average pairwise distance between units in each cluster. The second term is the average distance between all pairs of clusters. Since one usually wants to minimize the former term and maximize the latter, the value of *i* for which $g_1(i)$ has a minimum should be selected as the number of clusters.

7) Davies-Bouldin statistic.

This index proposed by Davies and Bouldin (1979) and discussed by Dubes (1982) is a compromise between compactness and isolation of the clusters. The formula is

$$D_2(i) = \left(\frac{1}{i}\right)\sum_{e=1}^{i} Max\left\{Q_{e,g}\right\}_{g=1,2,\ldots,k;\ e\neq g}; \qquad Q_{e,g} = \frac{\alpha_e/n_e + \alpha_g/n_e}{M(\mu_e,\mu_g)} \tag{60}$$

Where

$$\alpha_e = \sqrt{\psi_e},\ e = 1,2,\ldots,k;\ r = 1,2,\ldots,n-1,\ s = r+1,r+2,\ldots,n$$

Dubes (1982) adopted the following decision rule: if the sequence of $D_2(i)$ have a strong minimum at *i* then *k=i* is the true number of clusters. If the minimum is at *i=2* then *k=2* only if there is a significant drop in the values of (60).

8) Dubes index.

Dubes (1986) proposed a statistic which views the clustering solution as a model for the data.

$$\begin{gathered} D_3(i) = \frac{r - M_p M_c}{\sigma_p \sigma_c}; \qquad r = N^{-1}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}\sqrt{M(X_r,X_s)M(\mu_{\gamma_r},\mu_{\gamma s})}; \\ \sigma_p^2 = N^{-1}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}M(X_r,X_s) - M_p^2; \qquad \sigma_c^2 = N^{-1}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}M(\mu_{\gamma_r},\mu_{\gamma s}) - M_c^2; \\ M_p = N^{-1}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}\sqrt{M(X_r,X_s)}; \quad M_c = N^{-1}\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}M(\mu_{\gamma_r},\mu_{\gamma s}) \end{gathered} \tag{61}$$

The intuitive notion behind (61) is that the centroid represents the "true" positions of the entities so distances between centroids estimate the "true" distances between patterns. The statistic $D_3(i)$ tend to increase monotonically as the number of clusters increases so the best number of clusters is indicated by a significant knee in the curve of $D_3(i)$ as $i$ varies between $k_1$ and $k_2$.

9) Internal homogeneity of the clusters

$$\tau_1(i) = \frac{\underset{1\le j\le k}{Min}\left\{\rho_j \middle| n_j > 1\right\}}{\underset{1\le j\le k}{Max}\left\{\rho_j\right\}}, \qquad \rho_j = \underset{\gamma_r = j}{Max}\left\{M\left(X_r, \mu_j\right)\right\} \tag{62}$$

$\rho_j$ is the radius of cluster j, that is, the distance between a centroid and the farthest entity belonging to the cluster. If a clustering includes a singleton this is not considered because (61) would be always zero. We have: $0 \le \tau_1(i) \le 1$ for each i; $\tau_1(i) \to 1$ if the data set tends to form a single cluster and $\tau_2(i) \to 0$ if the best cluster solutions tends to the disjoint partition.

The index (62) is based on the assumption that cohesion is at its maximum level for a conjoint cluster (in this case $\tau_1(i)=1$) and, that any subdivision decreases general cohesion. As the number of cluster increases, the same groups become more homogeneous in their interior thus reducing the cluster radii. If the data set has i=k significantly compact and isolated clusters, its partition into $i=k+1$ clusters gives rise to a small reduction of $\tau_1(i)$ meaning that the additional information is not relevant. Conversely, if the number of clusters is stopped at $i=k-1$ then there is a less cohesion because heterogeneous entities are forcedly combined together. It follows that a relative minimum in the graph is a good candidate value for k; if, however, the clusters are widely separated, then the correct value of g would more probably be indicated by the peak point which almost always follows the minimum point. This is presumedly due to the sharp decrease in the maximal cluster radius at the denominator of $\tau_1(i)$ which, in such cases, undergoes a greater reduction than the numerator. A monotonic decreasing sequence is indicative that no cluster structure exists.

10) Overall separation between clusters.

$$\tau_2(i) = \frac{\sum_{j=1}^{k} n_j}{i}, \qquad \eta_j = \sum_{1\le r\le n;\gamma_r \ne j} M\left(X_r, \mu_j\right) \tag{63}$$

$\eta_j$ is the isolation coefficient of cluster $j$; that is, the distance between the centroid of the cluster and the nearest entity non belonging to the cluster (called the neighbor of the cluster). The

isolation coefficient is properly indeterminate for the conjoint cluster. According to (63), the isolation of a cluster is increases as the distance between the centroid and its neighbor is larger. As the number of clusters increases, global isolation intensifies because of the expanding disjunction of the cluster centroids. If the true number of clusters is *i=k* and the data set is divided into *i=k+1* groups, then the degree of their separation is reduced since the splitting of a homogeneous cluster yields smaller isolation coefficients. On the other hand, if the data are forcedly partitioned into *i=k-1* groups, then the inevitable lumping together of two or more of the nearest clusters (or portions of them) also produces a decrease in $\tau_2(i)$ because the centroids are now at a shorter distance from their neighboring entity. Consequently, a local maximum in (63) should be a reasonable estimate for the true number of clusters. If the values decrease monotonically, then *k=2* is likely to be the best choice for the number of clusters; if the values increase monotonically, then the data have no cluster structure or, rather, this cannot be uncovered by **Detclus**

11) Cubic clustering Criterion

Sarle (1983) used extensive simulation do develop a statistic which can be used for estimating the number of clusters. The formula is

$$C^3 = \left\{ Ln\left(\frac{1-E}{1-R^2}\right)\right\} \frac{\sqrt{\frac{nm^*}{2}}}{(0.001+E)^{1.2}}; \qquad R^2 = \frac{Trace(B)}{Trace(T)} = 1 - \frac{Trace(W)}{Trace(T)} \tag{64}$$

The formula of $R^2$ has the usual interpretation of the proportion of variance accounted for by the clusters. The properties of $R^2$ as a methods for determining the number of clusters are similar to *Min{|W(γ)|}*. However a plot of $R^2$ against *k* could be useful.

$$E = 1 - \left[ \frac{\sum_{j=1}^{m^*} \frac{1}{\left(n+u_j\right)} + \sum_{j=m^*+1}^{m} \frac{u_j^2}{\left(n+u_j\right)}}{\sum_{j=1}^{m} u_j^2} \right] \left[ \frac{(n-i)^2}{n} \right] \left( \frac{n+4}{n} \right)$$

$$m^* = \text{ largest integer less than m such that } u_{m^*} \geq 1 \tag{65}$$

$$u_j = \frac{s_j}{c}, \quad c = \left(\frac{v}{i}\right)^{\frac{1}{m^*}}, \quad v = \prod_{t=1}^{m^*} s_t, \; s_t = \sqrt{\lambda_{(t)}}$$

with $\lambda_{(t)}$ is the t-th eigenvalue of $(n-1)^{-1}\boldsymbol{T}$ arranged in decreasing order. The index $C^3$ is computed under assumption that the variables are uncorrelated (*e.g.* clustering of principal component scores). The following guidelines have been established.

1) Peaks of $C^3$>3 indicate good number of clusters.
2) Values of $C^3$ > indicate a likely good number of clusters
3) Values between 0 and 2 indicate potential clusters, but they should be taken with caution.
4) Large negative values may indicate outliers.
5) If $C^3$ continues to increase with the number of clusters, it may be an indication of subclusters in on or more clusters.

It goes without saying that these kind of checks are very heuristic and, perhaps, hazardous. As it is lucidly evidenced by Gordon (1981, p. 126), it seems unlikely that the indices such as those presented in this section and will find widespread acceptance in a strict hypothesis-testing sense. Their merit lies in the fact that can be easily computed and displayed, are clearly interpretable and require a minimum of human interaction allowing us to retain control over the classification process.

### 4.2.3 Experiments

This rather alarming number of rules is actually only a fraction of the statistics that have been devised so far; many methods were omitted from our discussion (for example the measures which depend too strongly upon the algorithm that generates the clustering: hierarchical procedures or iterative partitioning scheme with adaptive metrics) or which based on the stationary Poisson point process.

Why are there so many stopping rules? The answer is that each method has its own advantages and disadvantages so that it outperforms the others on some specific characteristics of the data set. Unfortunately, there is no known "best" stopping rule. There are many good methods depending on what is to be clustered, on what initial partition for what criterion. It is a good idea to learn the characteristics of each stopping rule, so that an intelligent choice can be made for particular applications. To this end, same of the above mentioned indices were applied to 24 well known data sets. The C-index, Klastorin, internal homogeneity index, and the Dubes index were not included in **DetClus** because their results, for the time being, do not legitimate the burden of computations.

It is well known that the behavior of the stopping rules depends on the quality of the final partition, which, in turn, depends on the starting partition. For this study , it has been preferred to try all the initialization methods and then keep the values of the rules corresponding to the lowest value of the criterion obtained across the runs. More specifically, the results reported in the tables are the best results attained for each value of *k* for all the initialization methods involving the algorithm TGBI (including a swapping phase as option 1). A combination of several

initialization methods is able to find a solution which is seldom very far from the best solution also for the weakly clustered data.

Comparisons of the stopping rules are usually carried out by Monte Carlo studies on idealized data. Another validation technique is to perturb the data, recluster them and compare the new and original clusterings, repeating the procedures using different degrees of contamination. Milligan (1996) observed that a test on an empirical data set is a sample size of one; the same is true for any trial in a Monte Carlo analysis since each non hierarchical clustering is a story in itself and cannot be confused with the computation of a statistics (each data set is unique from this particular point of view and should be examined separately). In this sense, replications of the same ideal or real situation with slight random or systematic changes in the entities are hardly useful in characterizing the behavior of the indices.

The 26 experiments can be subdivided into two types of situation.

*Strongly clustered (data set a-h)*

The variable space shows *k* high density regions entirely surrounded by empty space which can easily be detected by any technique worth of use. Any attempt to classify the entities in a number of clusters lower than or greater than *k* would create fictitious clusters. Measures of clusterability of the various alternative solutions would be very different and minor changes in the data, such as addition or removal of some observations, would be likely to lead to very similar clusterings.

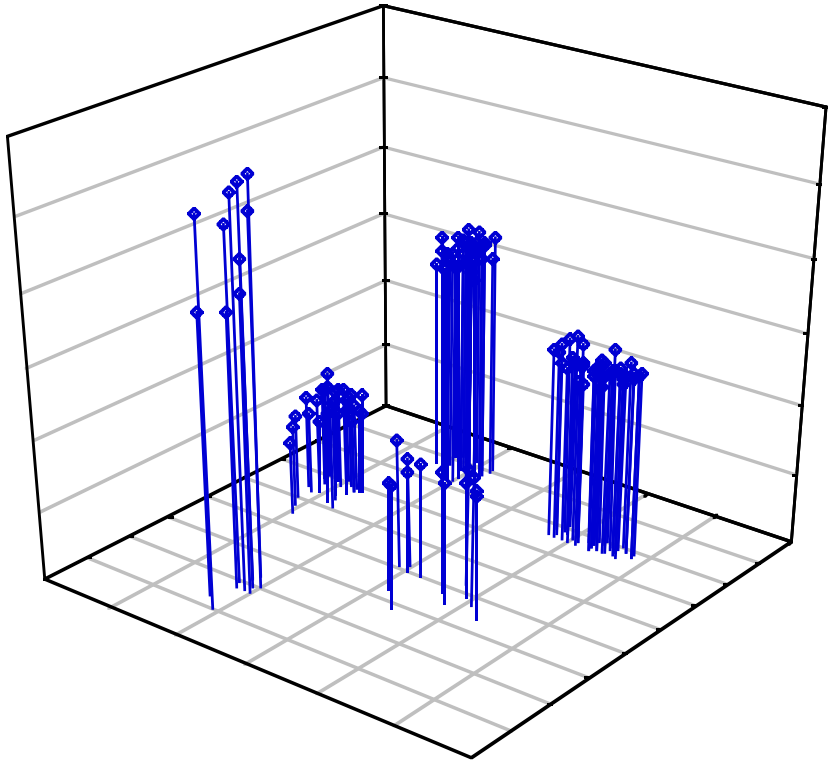
*poorly separated clustered (i-z)*

A cluster structure does exist, but the inherent variability of the data combined with the chaotic manifestations of chance has generated an anomalous sample. Such a sample includes not only entities belonging to a true group, but other erratic entities, isolated or hybrid, that will be inexorably recognized as "natural" clusters, this making difficult to determine the real values of *k*. A certain degree of misclassification is not surprising given the overlap of clusters and multiple local minima should be expected. For these data sets the user should opt for the most persistent solutions.

The range of the number of cluster is 2 to *k+4* where *k* is the known or the most indicated value for the number of clusters.

a) Artificial data set

This is a concocted example consisting of 140 five-dimensional points making up $k=6$ natural groups including 15, 25, 5, 40, 45, 10 entities. Each clusters is a region surrounding a local centroid. In this case the data set has both distinct and homogenous clusters that can easily be detected by any technique worthy of use.



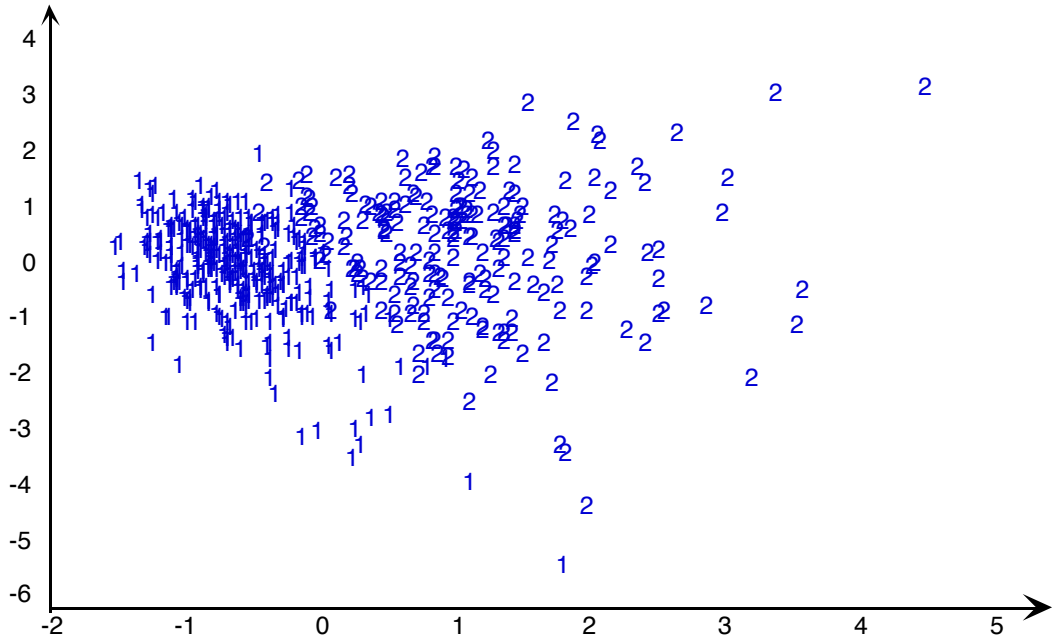
| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 1.10521 | 34.0 | 125.6 | 0.04421 | 0.9748 | 99.81 | 2.20 | 9.70 |
| 3 | 0.10151 | 41.6 | 131.5 | 0.00914 | 0.9612 | 6.68 | 1.36 | 10.07 |
| 4 | 0.01100 | 56.3 | 115.6 | 0.00176 | 0.9533 | 3.73 | 1.15 | 10.32 |
| 5 | 0.00125 | 48.9 | 297.2 | 0.00031 | 0.9257 | 15.00 | 0.90 | 10.33 |
| 6 | 0.00013 | 85.5 | 729.9 | 0.00005 | 0.9340 | 8.84 | 1.04 | 10.41 |
| 7 | 0.00004 | 184.1 | 752.4 | 0.00002 | 0.9165 | 11.45 | 0.89 | 10.52 |
| 8 | 0.00002 | 170.3 | 719.9 | 0.00001 | 0.9077 | 7.63 | 0.61 | 10.60 |
| 9 | 0.00001 | 155.1 | 710.4 | 0.00001 | 0.7773 | 2.97 | 0.57 | 10.66 |
| 10 | 0.00001 | 147.5 | 697.3 | 0.00001 | 0.7670 | 6.90 | 0.47 | 10.71 |

The Calinski index and CH suggest $k=7$ although there is a suspect discontinuity at $k=6$. The other indices have a peak or a valley at $k=6$ thus recovering the true number o clusters. The $C^3$ is not applicable to this data set. The results are expected since almost any stopping rule performs well when the clusters are sufficiently well separated.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 50 | 90 | | | | | | | | |
| 3 | 10 | 40 | 90 | | | | | | | |
| 4 | 10 | 5 | 85 | 40 | | | | | | |
| 5 | 50 | 10 | 25 | 15 | 40 | | | | | |
| 6 | 45 | 40 | 5 | 15 | 25 | 10 | | | | |
| 7 | 5 | 14 | 11 | 15 | 40 | 45 | 10 | | | |
| 8 | 5 | 11 | 4 | 40 | 14 | 15 | 45 | 6 | | |
| 9 | 6 | 16 | 15 | 4 | 11 | 14 | 29 | 5 | 40 | |
| 10 | 6 | 5 | 17 | 11 | 40 | 8 | 7 | 4 | 28 | 14 |

b) Wisconsin breast cancer data.

Each entity has 30 real-valued input features and an associated class labels (B=benign and M=malignant). The total number of entities are *n=569* (357B and 212M). A principal component analysis suggests that the first *m=25* factors (explaining 99.2% of total variation) are informative enough as to discriminate between benign and malignant cancers. The following figure represents the data set in the space of the first two principal components.



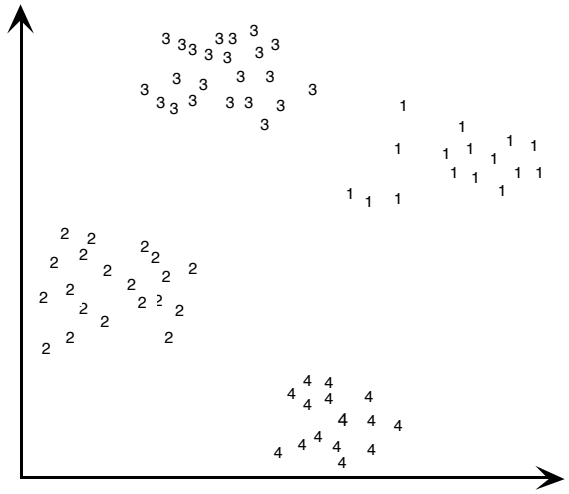

For *k=2* the recovery rate is 100%: $R_{HA}=1$ and *Q=1*. The correct number of cluster is indicated by all the indices

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.000628 | 23.6 | 31.7 | 0.0000 | 1.000 | 10289.97 | 1'185.37 | 14.30 |
| 3 | 0.000096 | 22.5 | 27.4 | 0.0000 | 0.985 | 0.25 | 404.10 | 13.99 |
| 4 | 0.000026 | 21.6 | 26.0 | 0.0000 | 0.958 | 0.13 | 1.60 | 13.81 |
| 5 | 0.000008 | 19.2 | 23.2 | 0.0000 | 0.781 | 0.13 | 1.10 | 13.81 |
| 6 | 0.000003 | 17.5 | 21.1 | 0.0000 | 0.721 | 0.13 | 0.99 | 13.79 |

| k | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 212 | 357 | | | | |
| 3 | 210 | 357 | 2 | | | |
| 4 | 210 | 355 | 2 | 2 | | |
| 5 | 175 | 355 | 2 | 2 | 35 | |
| 6 | 2 | 80 | 355 | 15 | 115 | 2 |

c) Ruspini data set. (Kaufman and Rousseeuw, 1990, p. 100).

This is a standard example consisting of 75 two-dimensional points making up *k=4* natural groups including 23, 20, 17, 15 entities. Actually the these data are different form the original data used by Ruspini (Rasson e Kukbushishi,1994, p. 191). Examination of the following figure shows that ***DetClus*** has a perfect recovery with this "ideal" structure.



Of course, it is reassuring to find that in this favorable situation the algorithm converged to the same partition yielding a correct reconstruction of the original classification.

The Ruspini data set has a crucial points at *k= 4* which is detected by almost all the applicable indices. It must be noted that CH, D1, and $\tau_2$ indicate that also *k=2* is a crucial point for *Min{|W(γ)|}* for this data set;

| k | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 35 | 40 | | | | | | |
| 3 | 27 | 35 | 13 | | | | | |
| 4 | 15 | 17 | 20 | 23 | | | | |
| 5 | 14 | 15 | 23 | 20 | 3 | | | |
| 6 | 23 | 15 | 8 | 14 | 12 | 3 | | |
| 7 | 15 | 20 | 8 | 13 | 4 | 12 | 3 | |
| 8 | 10 | 15 | 10 | 3 | 6 | 13 | 8 | 10 |

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 5.18675 | 65.8 | 126.7 | 0.20747 | 0.9433 | 32.388 | 0.536 | 8.287 |
| 3 | 1.99925 | 75.9 | 110.6 | 0.17993 | 0.8667 | 27.967 | 0.330 | 8.593 |
| 4 | 0.33925 | 323.4 | 425.3 | 0.05428 | 0.9152 | 23.299 | 0.383 | 8.669 |
| 5 | 0.20237 | 299.2 | 403.9 | 0.05059 | 0.8949 | 2.816 | 0.316 | 8.869 |
| 6 | 0.13242 | 256.2 | 373.8 | 0.04767 | 0.8307 | 4.176 | 0.230 | 9.020 |
| 7 | 0.09377 | 242.9 | 362.4 | 0.04595 | 0.8015 | 3.451 | 0.204 | 9.137 |
| 8 | 0.07113 | 345.1 | 367.5 | 0.04552 | 0.7272 | 5.130 | 0.176 | 9.230 |

d) Multiuniform data set.

This is an artificial data set containing four three-dimensional clusters compact and well separated simulated from the uniform distribution. The mean vectors are $\mu_i = d_i \boldsymbol{c}_m \quad i=1,2,...,k$ where $\boldsymbol{c}_m$ is a vector (mx1) of "1". The dependency between the variables of the clusters is specified by the variance-covariance matrix $\Sigma=(\sigma_{ij})$ where $\sigma_{ii}=9$, $\sigma_{ij}=3$ for $i \neq j$ which ensures an adequate cohesion for each cluster. The equal variance, equal covariance structure has been discussed by Wilks in the context of a two-way analysis of variance of the matrix of repeated measurements. It tacitly assumes that the variability of each cluster is described by a dominant factor accounting for $(100/m)[1+(m-1)/3]$ percent of the total variance in the cluster; the remaining variation is equally attributable to the other $(m-1)$ factors. The loadings of the dominant factor have both the same sign and the same absolute value: $m^{-0.5}$. Morrison (1967, 244-245) notes that this dimension is an average factor having an equiangular orientation in the midst of the axes of the original variables). The remaining components are bipolar factors because of the orthogonality with the first. The Mahalanobis distance between the centroids $\mu_i$ and $\mu_j$ is proportional to $|d_i - d_j|/2$ allowing for a check wether the centroids provide a satisfactory separation between clusters.

The following values were used for the simulations: $d_1=74, d_2=62, d_3=49, d_4=35, d_5=20$ which determine a sufficient isolation of the clusters and prevent atypical entities such as outliers which are unduly emphasized by the Mahalanobis metric and borderline entities which are difficult to classify in a hard clustering context. The cardinalities are n1=15, $n_2=25$, $n_3=35$, $n_4=45$, $n_5=55$. From a geometrical point of view, the clusters tend to take the form of a hyper-rectangle which calls in to question the performance of k-means algorithms based on $Min\{|\boldsymbol{W}|\}$ which is oriented to finding hyper-ellipsoidal clusters.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | $\tau 2$ | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 21.20425 | 26.2 | 6.1 | 0.84817 | 0.5163 | 4.245 | 0.041 | 12.380 |
| 3 | 9.65254 | 15.4 | 661.8 | 0.86873 | 0.4756 | 7.011 | 0.046 | 9.572 |
| 4 | 3.44757 | 11.5 | 1068.0 | 0.55161 | 0.5858 | 3.858 | 0.074 | 9.776 |
| 5 | 1.06443 | 8.9 | 1588.6 | 0.26611 | 0.7837 | 7.034 | 0.157 | 9.929 |
| 6 | 0.75165 | 9.5 | 1308.0 | 0.27060 | 0.6706 | 2.500 | 0.119 | 10.052 |
| 7 | 0.51641 | 9.5 | 1124.4 | 0.25304 | 0.5336 | 4.238 | 0.081 | 10.140 |
| 8 | 0.37074 | 10.5 | 981.3 | 0.23727 | 0.4517 | 3.381 | 0.057 | 10.205 |
| 9 | 0.23845 | 9.1 | 906.6 | 0.19315 | 0.5176 | 3.334 | 0.072 | 10.256 |

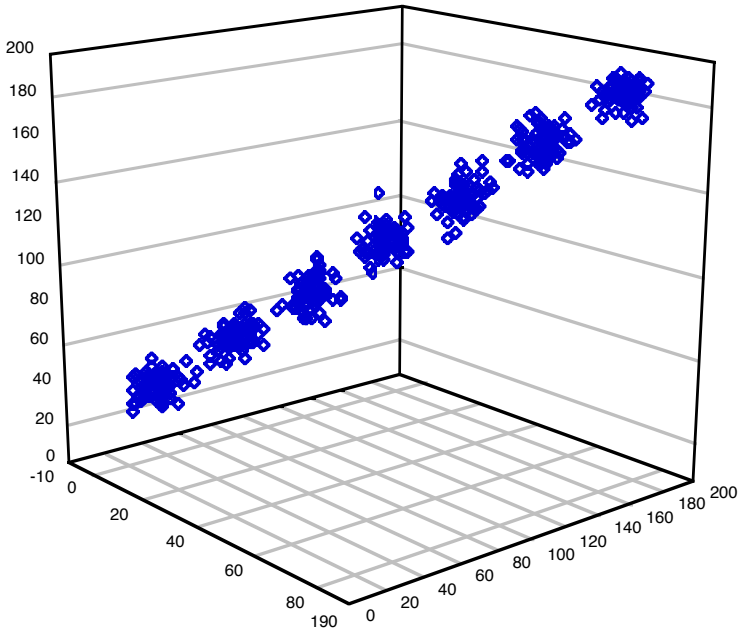
All the applicable indices have a crucial point (a local minimum or a local maximum) at $k=5$.

.

| k | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 89 | 86 | | | | | | |
| 3 | 56 | 41 | 78 | | | | | |
| 4 | 55 | 45 | 35 | 40 | | | | |
| 5 | 35 | 15 | 25 | 55 | 45 | | | |
| 6 | 19 | 35 | 26 | 55 | 25 | 15 | | |
| 7 | 15 | 21 | 20 | 25 | 34 | 35 | 25 | |
| 8 | 25 | 19 | 26 | 20 | 25 | 15 | 16 | 29 |
| 9 | 18 | 15 | 18 | 17 | 20 | 25 | 25 | 30 | 7 |

e) Unequal variances.

This data set has seven cluster each of which includes 70 entities generated according seven 8-dimensional normal distributions having the following parameters




$$\mu_i = 25*(i-1)\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} + 5\begin{bmatrix}1\\2\\\vdots\\8\end{bmatrix};$$

$$\Sigma_i = \begin{bmatrix}26 & 24 & \cdots & 24\\ 24 & 30 & \cdots & 24\\ \vdots & \vdots & & \vdots\\ 24 & 24 & \cdots & 40\end{bmatrix}$$

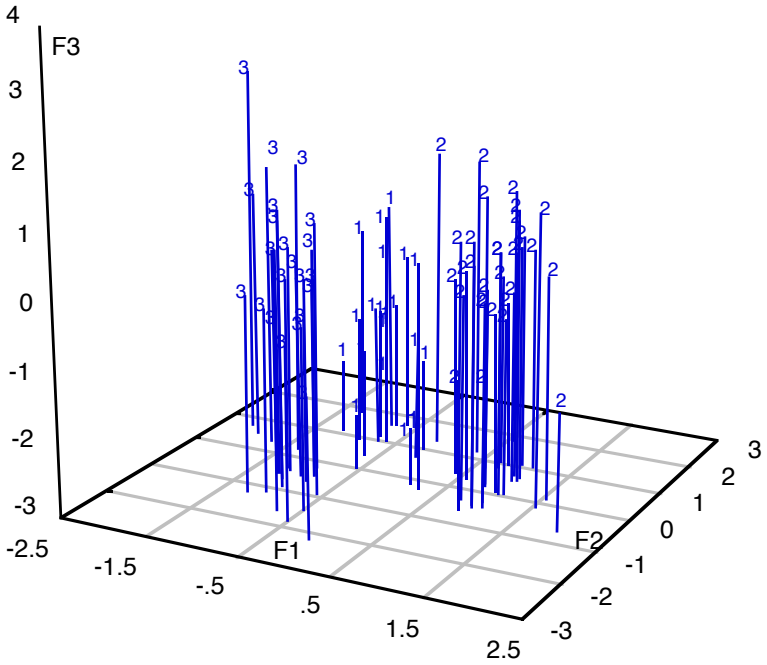
The intracluster covariance matrix has homogeneous correlations and distinct variances. This structure has been applied to model phenomenon that are permutation invariant, as for example, in the case of *m* equivalent psychological tests (Meza and Olkin, 1993). Since the data set is strongly clustered, the number of cluster underlying the data *(k= 7)* which is suggested by all the indices.

| k | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 242 | 248 | | | | | | | | | |
| 3 | 141 | 186 | 163 | | | | | | | | |
| 4 | 101 | 109 | 140 | 140 | | | | | | | |
| 5 | 125 | 89 | 84 | 95 | 97 | | | | | | |
| 6 | 70 | 75 | 135 | 70 | 70 | 70 | | | | | |
| 7 | 70 | 70 | 70 | 70 | 70 | 70 | 70 | | | | |
| 8 | 28 | 70 | 70 | 70 | 70 | 42 | 70 | 70 | | | |
| 9 | 41 | 70 | 35 | 70 | 29 | 70 | 70 | 70 | 35 | | |
| 10 | 35 | 36 | 70 | 45 | 70 | 34 | 70 | 35 | 70 | 25 | |
| 11 | 37 | 46 | 37 | 33 | 38 | 54 | 64 | 50 | 46 | 49 | 36 |

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 22.74 | 52.2 | 1291 | 0.910 | 0.428 | 1.78 | 0.01 | 10.52 |
| 3 | 9.78 | 31.5 | 1480 | 0.880 | 0.380 | 1.42 | 0.01 | 10.63 |
| 4 | 5.01 | 24.7 | 2144 | 0.802 | 0.378 | 1.40 | 0.02 | 10.71 |
| 5 | 3.22 | 19.0 | 1735 | 0.805 | 0.362 | 1.36 | 0.01 | 10.81 |
| 6 | 1.85 | 15.0 | 3148 | 0.665 | 0.446 | 1.29 | 0.02 | 10.84 |
| 7 | 0.31 | 12.9 | 6257 | 0.152 | 0.789 | 5.10 | 0.09 | 10.87 |
| 8 | 0.26 | 12.9 | 5444 | 0.168 | 0.710 | 1.17 | 0.07 | 10.90 |
| 9 | 0.21 | 13.0 | 4835 | 0.174 | 0.639 | 0.95 | 0.05 | 10.92 |
| 10 | 0.18 | 13.2 | 4399 | 0.180 | 0.566 | 0.87 | 0.04 | 10.94 |
| 11 | 0.71 | 18.4 | 841 | 0.858 | 0.266 | 0.92 | 0.01 | 11.00 |

f) Lubishew data set 1.

Discrimination in the genus *Chaetocnema*. Size: 74 observations, 6 variables (3 principal components), 3 clusters.



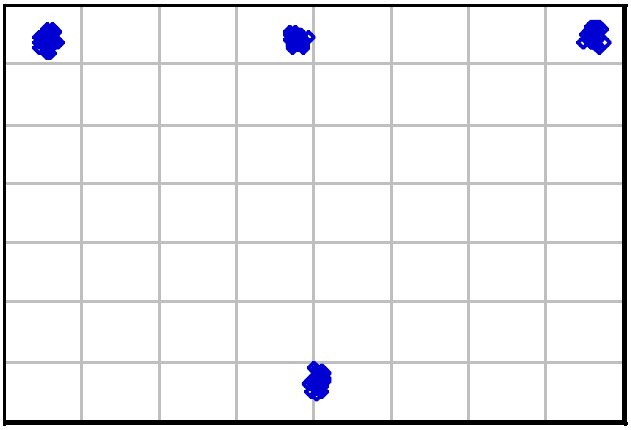

The data set has contains three natural clusters which can be clearly recovered. The value *k=3* is indicated by CH (which coincides with the Calinski), Marriott, C-index, $D_3$, and $\tau_2$. The S. coefficient, $D_1$, and $\tau_1$ indicate *k=2* or *k=3*. The index $C^3$ a minimum at *k=3*. It appears that both a local minimum and a local maximum are critical points for the cubic criterion..

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 10.5449 | 30.59 | 30.59 | 0.4218 | 0.838 | 8.382 | 0.207 | 9.266 |
| 3 | 1.73128 | 45.73 | 45.73 | 0.15581 | 0.830 | 7.395 | 0.244 | 9.125 |
| 4 | 0.93841 | 31.95 | 31.95 | 0.15015 | 0.723 | 4.561 | 0.212 | 9.439 |
| 5 | 0.65371 | 24.16 | 24.16 | 0.16343 | 0.658 | 5.413 | 0.155 | 9.678 |
| 6 | 0.39615 | 19.97 | 19.97 | 0.14261 | 0.626 | 2.192 | 0.067 | 9.847 |
| 7 | 0.26046 | 17.08 | 17.08 | 0.12763 | 0.615 | 2.308 | 0.079 | 9.979 |

| k | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 43 | 31 | | | | |
| 3 | 31 | 22 | 21 | | | |
| 4 | 22 | 15 | 16 | 21 | | |
| 5 | 6 | 15 | 16 | 15 | 22 | |
| 6 | 15 | 10 | 15 | 16 | 12 | 6 |
| 7 | 8 | 10 | 15 | 14 | 12 | 9 | 6 |

g) Wong data (Wong et al. 2001).

This is an artificial data set consisting of *n=200* ten-dimensional entities making up *k=4* natural groups each including 50 entities. The two-dimensional space spanned by the first two principal components (99.8% of the total variation) retains all the discriminant power among the entities. As can be seen, the five clusters obey the Van Rijsbergen's definition of perfect cluster and the Rao's string condition.



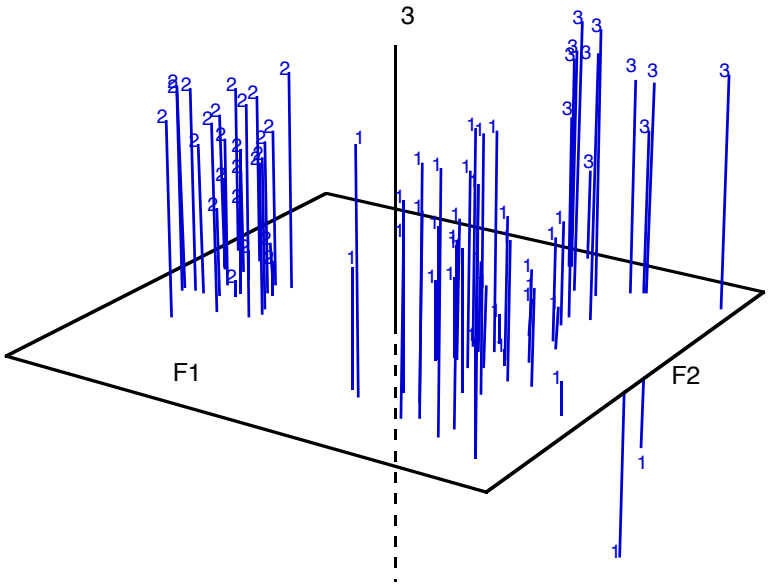
The true value of *k* has been detected by almost all the indices. Only $C^3$ gives a wrong indication. The two-cluster solution is considered a reasonable value for the number of clusters by the C-index, Silhouette coefficient, $D_1$ and $\tau_2$. This is not surprising if one considers the structure of the data and the "metric" definition of these indicators.

| k | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 150 | 50 | | | | | | |
| 3 | 100 | 50 | 50 | | | | | |
| 4 | 50 | 50 | 50 | 50 | | | | |
| 5 | 31 | 19 | 50 | 50 | 50 | | | |
| 6 | 22 | 28 | 22 | 50 | 28 | 50 | | |
| 7 | 26 | 13 | 11 | 50 | 29 | 50 | 21 | |
| 8 | 31 | 29 | 32 | 18 | 19 | 21 | 18 | 32 |

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.12114 | 198 | 198 | 0.00485 | 0.9991 | 317.70 | 20.30 | 9.79 |
| 3 | 0.02494 | 846 | 846 | 0.00224 | 0.8413 | 2.39 | 6.84 | 9.53 |
| 4 | 0.00005 | 82578 | 82578 | 0.00001 | 0.9992 | 951.49 | 22.89 | 9.63 |
| 5 | 0.00004 | 69619 | 69619 | 0.00001 | 0.8822 | 1.56 | 15.72 | 9.77 |
| 6 | 0.00003 | 66022 | 66022 | 0.00001 | 0.7686 | 1.18 | 8.85 | 9.87 |
| 7 | 0.00002 | 64399 | 64399 | 0.00001 | 0.6702 | 1.33 | 3.54 | 9.95 |
| 8 | 0.00002 | 51509 | 51509 | 0.00001 | 0.5951 | 1.42 | 0.01 | 10.02 |

h) Fossil specimens data set (Chernoff,1973).

Discrimination on nummulited specimens from Eocene yellow Limestone Formation of north-western Jamaica. Size: 87 observations, 6 variables (4 principal components), 3 clusters with cardinalities: 43, 13, 34.




The pseudo F statistics correctly spotted the true value of *k*. The value *k=3* is also suggested by the Marriott index, $D_1$, and $C^3$. The index $\tau_1$ gives *k=2* although there is an interesting strong decrease after *k=3*.

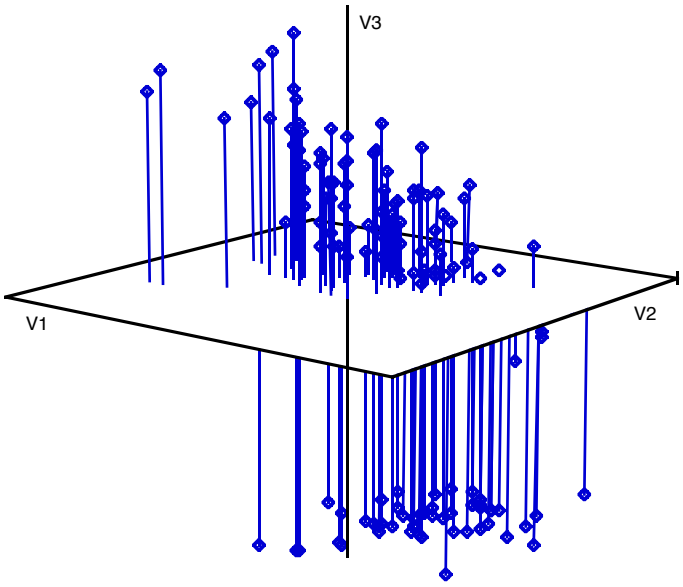
| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|
| 2 | 7.40866 | 25.6 | 25.6 | 0.29635 | 0.8603 | 3.98 | 0.31 | 9.93 |
| 3 | 1.31860 | 32.4 | 32.4 | 0.11867 | 0.8285 | 1.63 | 0.28 | 9.78 |
| 4 | 0.53943 | 30.9 | 30.9 | 0.08631 | 0.7199 | 2.26 | 0.22 | 9.93 |
| 5 | 0.31099 | 36.8 | 36.8 | 0.07775 | 0.7165 | 0.84 | 0.20 | 9.99 |
| 6 | 0.17675 | 30.5 | 30.5 | 0.06363 | 0.5778 | 1.10 | 0.12 | 10.16 |
| 7 | 0.11485 | 27.1 | 27.1 | 0.05628 | 0.5810 | 0.95 | 0.13 | 10.29 |

| k | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 53 | 34 | | | | | |
| 3 | 40 | 34 | 13 | | | | |
| 4 | 20 | 34 | 20 | 13 | | | |
| 5 | 6 | 34 | 13 | 13 | 21 | | |
| 6 | 12 | 11 | 13 | 20 | 23 | 8 | |
| 7 | 17 | 5 | 13 | 17 | 5 | 17 | 13 |

To see how various stopping rules differ, we must examine data set in which the clusters are much close together.

i) The Iris Plants.

This is perhaps the best known data set to be found in the cluster analysis literature. There are 4 measurements on 50 plants from each of 3 species of *Iris*: *setosa, versicolor, virginica*. The *Setosa* plants are linearly separable from the other. *Versicolor* is an hybrid of *Setosa* and *Virginica*, but much more similar to the latter; consequently, there is some overlap between the two species. ***DetClus*** provides a separation into 3 clusters: (50,0,0); (0,48,2), (0,1,49) which agrees with the findings of Friedman and Rubin (1967) and Maronna and Jacovkis (1974). Three undecided cases is probably the best possible results with this data set (see Richards, 1972).



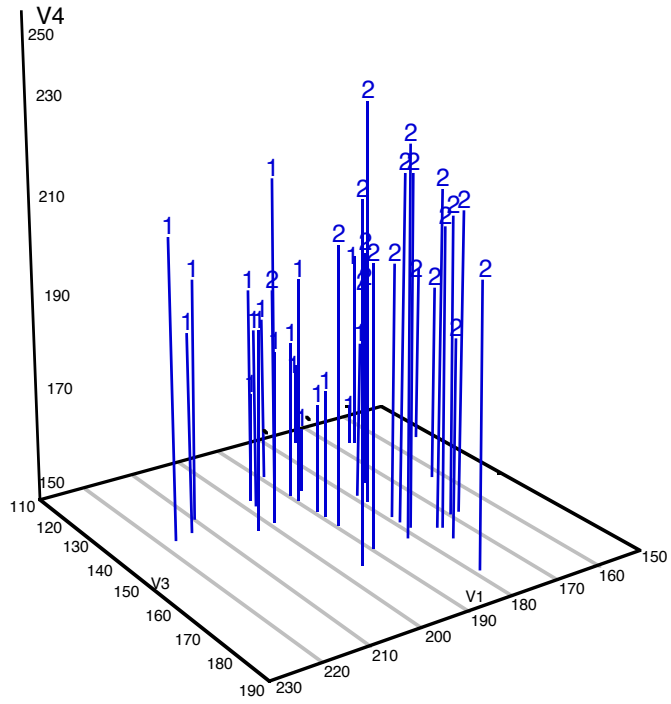

| k | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 100 | 50 | | | | | |
| 3 | 50 | 49 | 51 | | | | |
| 4 | 34 | 18 | 48 | 50 | | | |
| 5 | 50 | 12 | 29 | 27 | 32 | | |
| 6 | 50 | 28 | 19 | 23 | 18 | 12 | |
| 7 | 18 | 25 | 12 | 21 | 26 | 23 | 25 |

Most of the indices fail miserably yielding *k=2* as the correct number of clusters (although the Calinski index has a minimum at *k=3*). Only the Marriott index gives *k=3*. The table provides no sharp conclusions about the appropriate value of *k*. This is not a total surprise since, as it is well known, two categories of the data set cannot be separately completely by hyperplanes. Worthy of note here is that Dubes (1987) did non found a local minimum at *k=3* for index $D_2$ based on the Euclidean metric. This support the conjecture that the stopping rules behave differently under different clustering criterion.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 9.20049 | 43.46 | 502.82 | 0.36802 | 0.838 | 5.410 | 0.562 | 0.183 | 9.171 |
| 3 | 2.20397 | 31.89 | 483.08 | 0.19836 | 0.736 | 8.439 | 0.771 | 0.162 | 9.504 |
| 4 | 0.91958 | 37.67 | 364.24 | 0.14713 | 0.733 | 3.388 | 0.843 | 0.136 | 9.749 |
| 5 | 0.58803 | 35.17 | 413.82 | 0.14701 | 0.674 | 4.657 | 0.944 | 0.117 | 10.257 |
| 6 | 0.35936 | 29.12 | 398.03 | 0.12937 | 0.657 | 3.012 | 0.980 | 0.136 | 10.384 |
| 7 | 0.23523 | 28.55 | 400.97 | 0.11526 | 0.499 | 3.873 | 1.187 | 0.038 | 10.470 |

j) Lubishew data set 2 (Lubishew, 1962).

Two cryptic species of the flea beetles genus *Halticus* were separated (19 specimen of *H. oleracea* and 20 of *H. carduorum)* using 4 external characters.




| k | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 18 | 21 | | | | |
| 3 | 10 | 16 | 13 | | | |
| 4 | 10 | 12 | | 9 | | |
| 5 | 5 | 9 | | 8 | 8 | |
| 6 | 7 | 8 | | 5 | 4 | 7 |

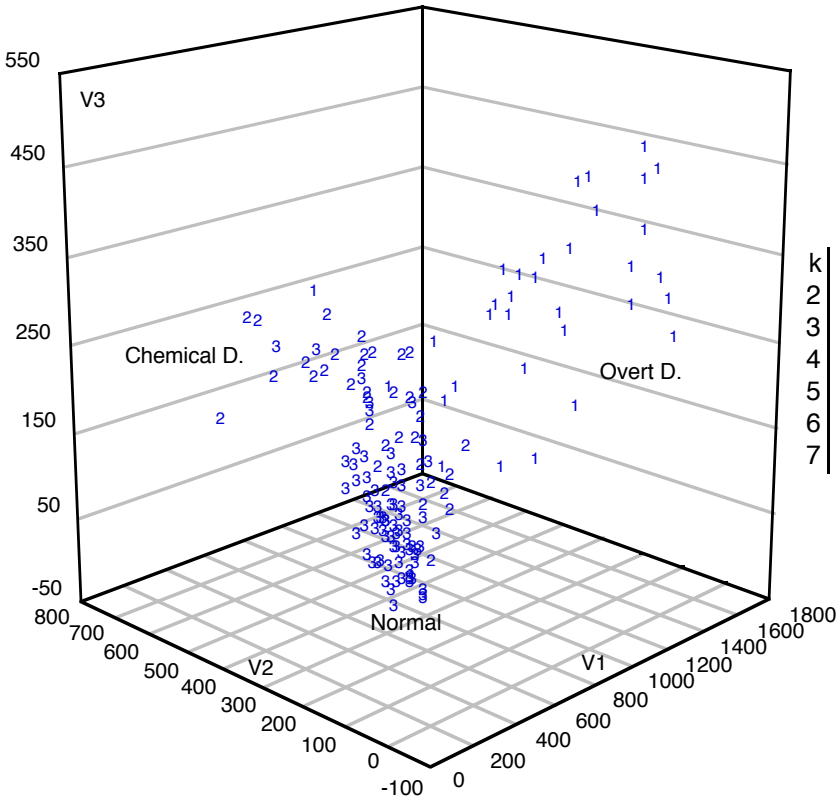
In this case the dispersion matrices are very different for the two types and it is doubtful whether it is even legitimate to suppose they have a common dispersion matrix. Nevertheless, **DetClus** retrieved almost exactly the two genies except for 2 *H. oleracea* and 1 *H. cardiourum*. The recovery rate of **DetClus** is fairly satisfactory and it may be reasonable expected that the program is capable to reconstruct the true underlying cluster structure (if a cluster structure is actually present) in the data set. A certain degree of misclassification is to be expected given the natural overlap of clusters in most real applications.

The appropriate value *k=2* was detected by Calinski, CH, $D_1$, and $D_2$. The silhouette coefficient suggests *k=3*. The index $\tau_2$ yield *k=4*. The Marriott index gives no clear indication.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 18.5855 | 9.46 | 15.31 | 0.74342 | 0.646 | 7.840 | 0.950 | 0.224 | 8.208 |
| 3 | 5.18231 | 6.21 | 4.52 | 0.46641 | 0.724 | 8.735 | 0.843 | 0.380 | 9.213 |
| 4 | 2.04037 | 6.11 | 1.88 | 0.32646 | 0.700 | 5.290 | 0.817 | 0.394 | 10.277 |
| 5 | 1.04495 | 6.91 | 9.08 | 0.26124 | 0.663 | 6.038 | 0.914 | 0.309 | 9.330 |
| 6 | 0.49442 | 9.01 | 15.23 | 0.17799 | 0.685 | 4.908 | 0.831 | 0.386 | 9.308 |

k) Chemical and overt Diabetes (Andrews and Herzberg,1985).

This data set consists of variables (insulin area, glucose area, and steady-state plasma glucose response) measured on *n=145* non obese adult subjects. The subjects were clinically classified as normal (76), Chemical diabetes (36) and overt diabetes (33). The clusters have various sizes and different non-ellipsoidal dispersion matrices.



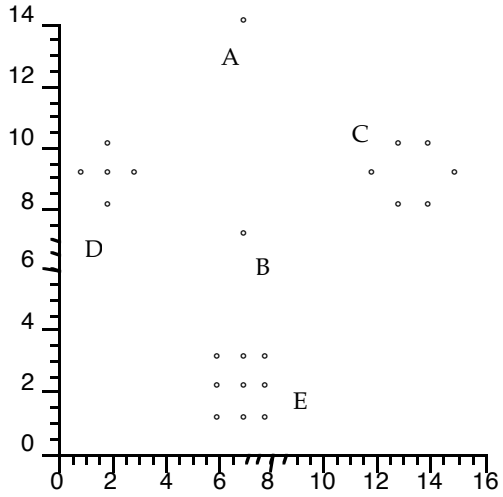

| k | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | 119 | 26 | | | | | |
| 3 | 45 | 80 | 20 | | | | |
| 4 | 14 | 12 | 79 | 40 | | | |
| 5 | 14 | 38 | 71 | 12 | 10 | | |
| 6 | 38 | 19 | 55 | 12 | 14 | 7 | |
| 7 | 27 | 15 | 19 | 12 | 7 | 39 | 26 |

Assuming homogeneous dispersion matrices **DetClus** found (72,7,1) (4, 29,12) (0,0,20) with an error rate of 16.6% which is aligned with those obtained by Banfield and Raftery (1993). The true number of clusters is produced by Calinski and $D_2$ and, perhaps, Marriott. It is worthy of note, however, that at *k=3*, the plot of CH index, silhouette coefficient, $D_1$ and $\tau_2$ have a critical point which has an opposite form (a peak instead of a valley or *vice versa*) of what expected for well structured data sets.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 16.21430 | 28.79 | 356.48 | 0.64857 | 0.773 | 2.384 | 0.840 | 0.113 | 9.17 |
| 3 | 4.36246 | 30.23 | 270.03 | 0.39262 | 0.592 | 3.850 | 1.050 | 0.064 | 9.87 |
| 4 | 1.28045 | 26.61 | 315.14 | 0.20487 | 0.608 | 3.554 | 0.964 | 0.102 | 10.11 |
| 5 | 0.56755 | 27.72 | 329.11 | 0.14189 | 0.573 | 1.755 | 0.982 | 0.100 | 10.28 |
| 6 | 0.29710 | 28.73 | 268.88 | 0.10696 | 0.512 | 2.419 | 1.063 | 0.090 | 10.41 |
| 7 | 0.18674 | 28.24 | 222.58 | 0.09150 | 0.481 | 2.813 | 1.097 | 0.081 | 10.51 |

l) Fuzzy data set (Kaufman and Rousseeaw, 1990, p. 144).

The data set contains 22 points characterized by two variables. Three main clusters and two intermediate entities can be visually distinguished: "A" is an isolated cluster and "B " is an hybrid entity. If k=3 any Iterative partitioning method would have to make a rather arbitrary choice as to whether to attach entity A to the cluster C or D. Also the assignment of entity B can be very difficult since its membership is spread out over the other clusters.




| k | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 15 | 7 | | | | | | | |
| 3 | 7 | 9 | 6 | | | | | | |
| 4 | 6 | 6 | 9 | 1 | | | | | |
| 5 | 6 | 9 | 5 | 1 | 1 | | | | |
| 6 | 6 | 3 | 6 | 1 | 1 | 5 | | | |
| 7 | 3 | 5 | 1 | 5 | 3 | 1 | 4 | | |
| 8 | 3 | 5 | 3 | 3 | 2 | 1 | 4 | 1 | |
| 9 | 3 | 3 | 2 | 3 | 3 | 1 | 3 | 3 | 1 |

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | $\tau 2$ | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 6.33243 | 17.62 | 22.54 | 0.2533 | 0.937 | 15.895 | 0.334 | 1.807 | 6.437 |
| 3 | 1.06528 | 48.75 | 52.60 | 0.0959 | 0.909 | 18.341 | 0.416 | 1.343 | 6.480 |
| 4 | 0.37924 | 76.91 | 78.74 | 0.0607 | 0.927 | 4.258 | 0.324 | 1.904 | 6.738 |
| 5 | 0.13443 | 110.90 | 112.66 | 0.0336 | 0.922 | 2.481 | 0.223 | 2.450 | 6.965 |
| 6 | 0.08402 | 104.35 | 103.69 | 0.0303 | 0.798 | 4.060 | 0.446 | 2.031 | 7.167 |
| 7 | 0.04133 | 98.65 | 101.68 | 0.0203 | 0.784 | 5.196 | 0.521 | 1.410 | 7.332 |
| 8 | 0.02505 | 86.7 | 90.5 | 0.0160 | 0.778 | 5.595 | 0.545 | 1.161 | 7.474 |
| 9 | 0.01604 | 71.4 | 77.9 | 0.0130 | 0.785 | 6.160 | 0.550 | 1.259 | 7.597 |

The choice *k=5* is indicated by Calinski, CH index, $D_1$, $D_2$, $\tau_2$. Marriott suggests *k=3*. The value *k=4* is a possibility for the Silhouette coefficient. It is evident that the determination of the number of cluster cannot be considered separately from the role assigned to the outlier A and the hybrid B. With *k=3* **DetClus** added A to cluster C and B to cluster D.

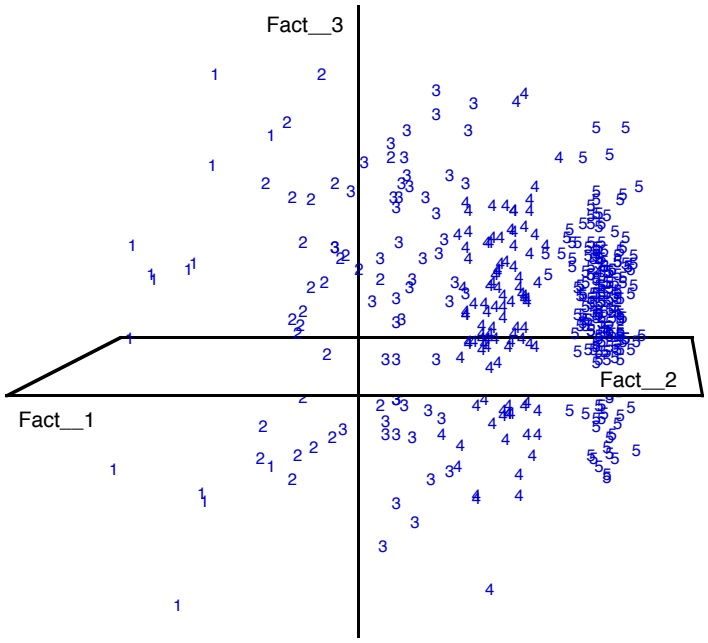
m) Clusters of unequal size and dispersion

In this data set there are five groups including respectively (15,30,60,120,240) entities generated according to *5* five-dimensional normal distributions having means and dispersion matrices

$$\mu_i = (-5+10i)\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}; \qquad \Sigma_i = \begin{bmatrix} 6i & 3i-1 & \cdots & 3i-1 \\ 3i-1 & 6i & \cdots & 3i-1 \\ \vdots & \vdots & & \vdots \\ 3i-1 & 3i-1 & \cdots & 6i \end{bmatrix}; \qquad i = 1,2,\ldots,k$$

Consequently, the clusters have different ellipsoidal shape, but the same orientation. The recovery rate of **DetClus** is satisfactory since only 23 entities were misclassified for *k=3*.

The CH index, Marriott, silhouette coefficient, $D_1$ and $\tau_2$ indicate *k=5* as the best classification of the data set. The Calinski index shows a decreasing sequence of values thus evidencing the inconsistency between the model and the data. The Davies-Bouldin statistics has a critical point at *k=6*. However, we should note that the Calinski index has relatively small increments after *k=5* an d that $D_2$ start to increase just for values of *k* greater than 5. All considered, there is sufficient evidence to select *k=5* as the true number of clusters.

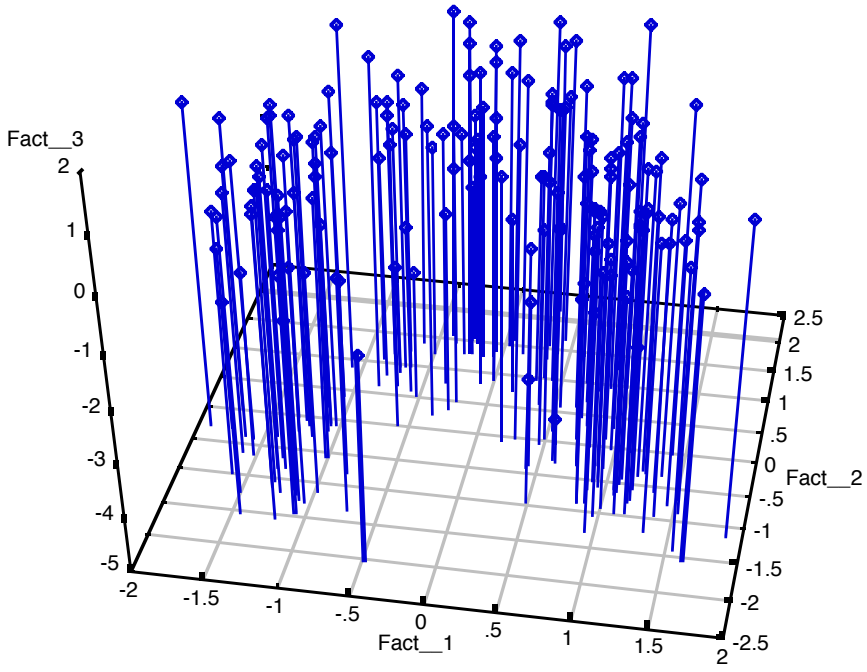



| k | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 346 | 119 | | | | | | |
| 3 | 152 | 63 | 250 | | | | | |
| 4 | 240 | 19 | 137 | 69 | | | | |
| 5 | 55 | 240 | 16 | 124 | 30 | | | |
| 6 | 113 | 16 | 65 | 107 | 128 | 36 | | |
| 7 | 114 | 17 | 47 | 125 | 35 | 56 | 71 | |
| 8 | 15 | 144 | 30 | 29 | 96 | 26 | 67 | 58 |
| 9 | 56 | 25 | 80 | 74 | 16 | 86 | 35 | 30 | 63 |

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 28.8238 | 76.85 | 629.7 | 1.1530 | 0.569 | 2.325 | 1.356 | 0.0101 | 10.752 |
| 3 | 11.0168 | 50.96 | 1182.4 | 0.9915 | 0.527 | 2.947 | 1.216 | 0.0133 | 10.654 |
| 4 | 5.5636 | 39.55 | 1327.2 | 0.8902 | 0.601 | 3.459 | 1.077 | 0.0197 | 10.738 |
| 5 | 3.3323 | 31.72 | 1361.6 | 0.8331 | 0.641 | 2.574 | 1.123 | 0.0231 | 10.802 |
| 6 | 2.3257 | 27.95 | 1013.6 | 0.8372 | 0.412 | 3.200 | 1.312 | 0.0198 | 10.869 |
| 7 | 1.5523 | 23.79 | 898.0 | 0.7606 | 0.434 | 3.707 | 1.267 | 0.0221 | 10.907 |
| 8 | 1.1800 | 28.09 | 883.6 | 0.7552 | 0.405 | 2.285 | 1.356 | 0.0188 | 10.927 |
| 9 | 0.9178 | 31.32 | 749.7 | 0.7434 | 0.372 | 3.080 | 1.392 | 0.0199 | 10.952 |

n) Wine recognition data. (Blake and Merz, 1998).

These data are the results of a chemical analysis of *n=178* wines grown in the same region in Italy but derived from *k=3* different cultivars. The analysis determined the quantities of *13* constituents found in each of the three type of wine. The class appears to be linearly separable with (59,71,48) members per class.

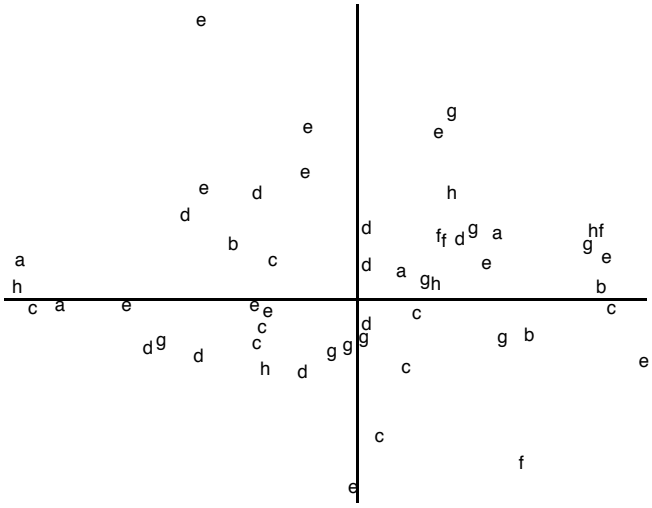



The true number of clusters was found by Calinski, CH, and Marriott. A local peak is found at *k=2* for the silhouette coefficient and $\tau_2$. The Dunn index and the Davies-Bouldin yield *k=5*. It must be noted that the final solution for *k=3* has a modified Rand index of $R_{HA}=0.9188$ with 4 entities misclassified. On the other hand, one of the best solutions for *k= 3* attained $R_{HA}=0.9651$ with only two entities misclassified, but the criterion value was 1.8856 denoting an inferior partition to that accepted by **DetClus**.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 11.3202 | 12.89 | 13.27 | 0.4528 | 0.575 | 11.766 | 1.175 | 0.1255 | 11.987 |
| 3 | 1.8583 | 13.27 | 205.59 | 0.1673 | 0.529 | 3.245 | 1.295 | 0.1021 | 10.156 |
| 4 | 0.6805 | 12.74 | 137.51 | 0.1089 | 0.496 | 3.135 | 1.537 | 0.0940 | 10.396 |
| 5 | 0.2730 | 11.21 | 109.70 | 0.0683 | 0.445 | 2.921 | 1.573 | 0.0769 | 10.515 |
| 6 | 0.1230 | 11.37 | 90.23 | 0.0443 | 0.457 | 2.996 | 1.493 | 0.1005 | 10.593 |
| 7 | 0.0579 | 11.11 | 125.37 | 0.0284 | 0.392 | 1.194 | 1.560 | 0.0983 | 10.532 |

| k | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 54 | 124 | | | | |
| 3 | 59 | 66 | 53 | | | |
| 4 | 36 | 34 | 49 | 59 | | |
| 5 | 56 | 24 | 27 | 37 | 34 | |
| 6 | 27 | 6 | 27 | 24 | 59 | 35 |
| 7 | 33 | 23 | 27 | 6 | 30 | 25 | 34 |

o) Painters data set (Davenport and Studdert-Kennedy, 1972).
In 1708 the French art critic R. De Piles published a book which contained aesthetic judgements on painters based on four subjective criteria: composition, drawing, color and expression. These judgements were expressed in form of scores attributed to each painter for each criterion. "Merit" scores were given complete for 54 painters divided into 8 schools. The figure shows the schools of R. De Piles in the space of the first two principal components (82.3% of variance explained).



Davenport and Studdert-Kennedy (1972) used cluster analysis to group painters with similar scores, but found little correspondence between clusters and schools. They considered *k=5* and *k=8* appropriate choices for the true number of natural clusters, but the results are not clear enough to conclude. In fact, the Calinski index, $D_1$ and $D_2$ support the former choice. The silhouette coefficient and $\tau_2$ suggest the latter. Marriott; CH and $C^3$ (note, however, that both The Marriott and $C^3$ decrease monotonically with k which is usually a sign that the data have a cluster structure.).

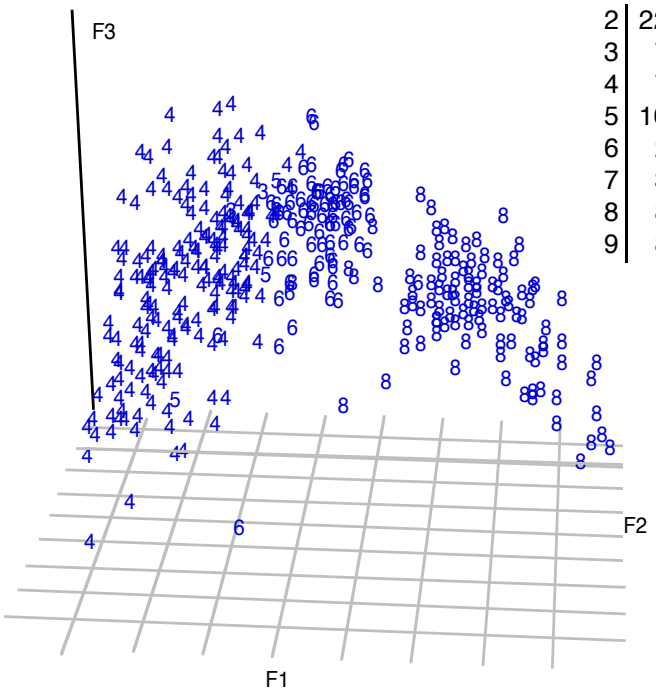
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 22 | 32 | | | | | | | |
| 3 | 21 | 21 | 12 | | | | | | |
| 4 | 11 | 10 | 19 | 14 | | | | | |
| 5 | 8 | 20 | 13 | 10 | 3 | | | | |
| 6 | 11 | 9 | 2 | 9 | 11 | 12 | | | |
| 7 | 2 | 7 | 10 | 6 | 6 | 11 | 12 | | |
| 8 | 5 | 11 | 5 | 2 | 10 | 8 | 9 | 4 | |
| 9 | 5 | 7 | 6 | 6 | 3 | 12 | 5 | 7 | 3 |
| 10 | 6 | 5 | 2 | 5 | 4 | 4 | 11 | 5 | 4 | 8 |

The solution *k=2* is the only finding for which the indices are concordant. Unfortunately, this is a structure which, though real is not necessarily of interest since a two-cluster solution is often generated merely by chance fluctuation. Only experts can decide wether it is interesting or a random aspects of the data.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 19.0158 | 13.20 | 21.08 | 0.7606 | 0.6540 | 7.977 | 0.962 | 0.118 | 8.813 |
| 3 | 7.6203 | 9.09 | 14.75 | 0.6858 | 0.6219 | 3.043 | 0.993 | 0.150 | 9.160 |
| 4 | 2.7279 | 9.89 | 12.05 | 0.4365 | 0.6372 | 6.252 | 0.913 | 0.186 | 9.427 |
| 5 | 1.6667 | 11.08 | 12.62 | 0.4167 | 0.6170 | 1.646 | 0.915 | 0.175 | 9.753 |
| 6 | 0.8580 | 9.35 | 12.82 | 0.3089 | 0.6050 | 6.091 | 0.909 | 0.221 | 9.881 |
| 7 | 0.5272 | 12.69 | 19.32 | 0.2583 | 0.6362 | 4.371 | 0.833 | 0.223 | 9.852 |
| 8 | 0.2863 | 6.79 | 9.46 | 0.1832 | 0.6662 | 2.898 | 0.833 | 0.270 | 10.188 |
| 9 | 0.1934 | 11.89 | 17.30 | 0.1566 | 0.6329 | 2.735 | 0.910 | 0.227 | 10.060 |
| 10 | 0.1200 | 6.78 | 11.03 | 0.1200 | 0.6619 | 3.208 | 0.911 | 0.258 | 10.240 |

p) Auto-Mpg Data (http://lib.stat.cmu.edu/datasets/).

The data concerns city-cycles fuel consumption in miles per gallon for n=392 cars. Only continuous variables were used for the clustering. In particular: mpg, displacement, horsepower, weight, acceleration. The number of cluster is unknown. In the space of the first three principal components (97.3% of total variation explained) the cars appears distinguishable by the number of cylinders (72.4%) of correct classification.



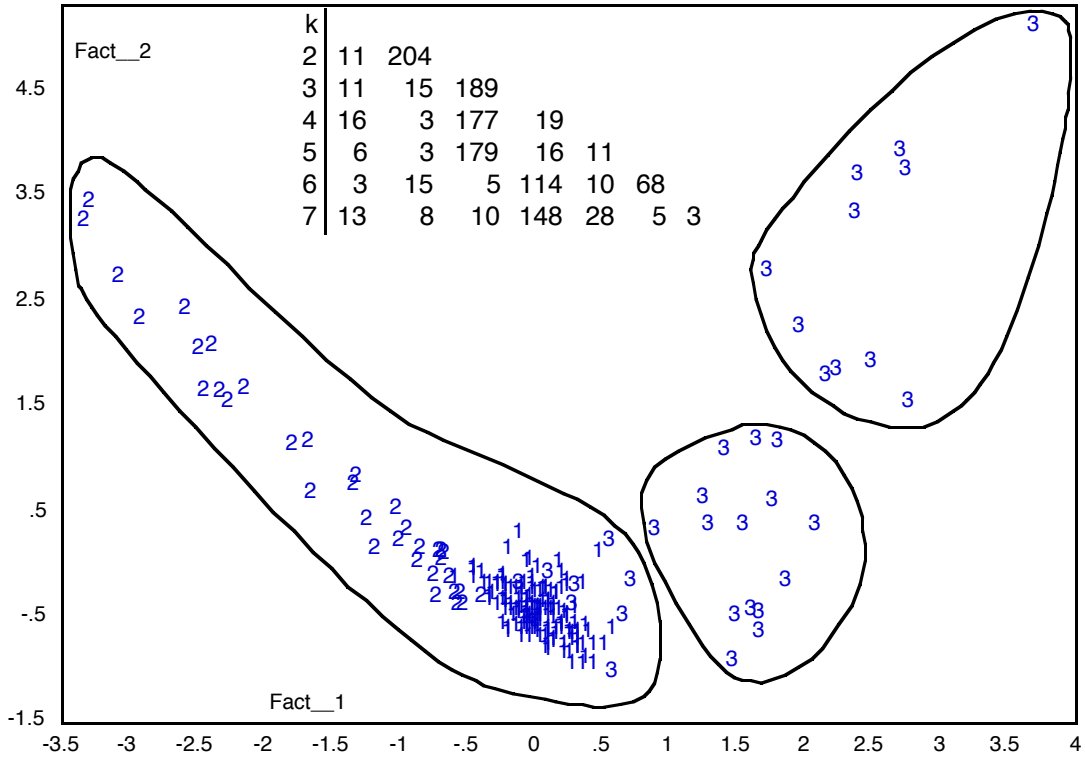

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 225 | 167 | | | | | | |
| 3 | 73 | 98 | 221 | | | | | |
| 4 | 72 | 125 | 98 | 97 | | | | |
| 5 | 100 | 76 | 23 | 72 | 121 | | | |
| 6 | 22 | 98 | 71 | 76 | 53 | 72 | | |
| 7 | 32 | 22 | 76 | 53 | 72 | 86 | 51 | |
| 8 | 82 | 71 | 17 | 5 | 33 | 77 | 50 | 57 |
| 9 | 85 | 21 | 59 | 15 | 53 | 51 | 71 | 5 32 |

The indication on the underlying number of clusters is discordant across the indices: Calinski, silhouette coefficient, $\tau_2$, $D_1$ indicate $k=2$. The index CH, $D_2$ advocate $k=3$. Only the Marriott index suggests $k=5$ which correspond to the five distinct number of cylinders revealing the true classification.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 13.141 | 82.0 | 701 | 0.526 | 0.717 | 2.98 | 0.87 | 0.027 | 10.72 |
| 3 | 5.041 | 67.4 | 757 | 0.454 | 0.618 | 1.13 | 0.67 | 0.019 | 10.80 |
| 4 | 2.170 | 70.1 | 617 | 0.347 | 0.491 | 1.40 | 0.72 | 0.013 | 10.90 |
| 5 | 1.034 | 69.5 | 474 | 0.259 | 0.523 | 1.28 | 0.82 | 0.017 | 10.96 |
| 6 | 0.655 | 66.1 | 340 | 0.236 | 0.462 | 1.21 | 0.86 | 0.014 | 11.02 |
| 7 | 0.466 | 66.0 | 294 | 0.228 | 0.460 | 1.00 | 0.79 | 0.015 | 11.04 |
| 8 | 0.331 | 66.3 | 244 | 0.212 | 0.467 | 0.37 | 0.83 | 0.020 | 11.06 |
| 9 | 0.242 | 59.5 | 214 | 0.196 | 0.456 | 0.63 | 0.70 | 0.021 | 11.07 |

q) Thyroid gland data (Blake and Merz, 1998).

Five laboratory tests are used to try to predict wether a patient's t thyroid to the class eu-, hypo- or hyper-thyroidism. The diagnosis (the class label) was based on a complete medical record, including anamnesis, scan etc. The class distribution is (150, 30, 35). The figure below shows the data in the plan of the first two principal components. It is evident that the hypothesis of homogeneous variance-covariance matrix across the groups is violated. The recovery rate of **DetClus** (defined by the lines) is unsatisfactory: the cases of hypothyroidism are completely missed and those of hyperthyroidism are splitted in two anomalous subgroups.

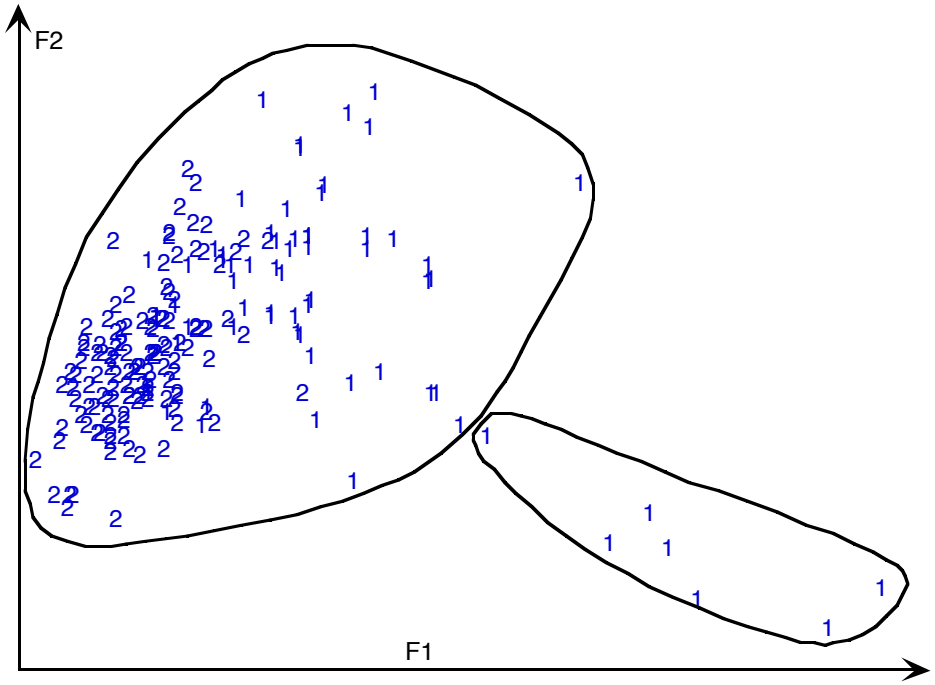



Nevertheless, the value *k=3* sounds reasonable for the Calinski index, CH, $D_2$ and $C^3$. Marriott, silhouette coefficient $D_1$ and $\tau_2$ suggest *k=5* which does not appear consistent with these data. It is worth of note the fact that the true number of clusters can be successfully detected even if the recovery rate is unsatisfactory.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 19.8149 | 40.68 | 67.21 | 0.7926 | 0.9008 | 89.368 | 0.743 | 0.138 | 10.981 |
| 3 | 5.3122 | 46.89 | 94.27 | 0.4781 | 0.8232 | 8.492 | 0.754 | 0.085 | 10.744 |
| 4 | 1.5686 | 45.12 | 88.28 | 0.2510 | 0.8177 | 2.690 | 0.733 | 0.188 | 10.839 |
| 5 | 0.4176 | 52.42 | 86.77 | 0.1044 | 0.8310 | 1.501 | 0.643 | 0.230 | 11.099 |
| 6 | 0.1800 | 62.04 | 119.57 | 0.0648 | 0.4905 | 3.146 | 0.894 | 0.193 | 11.052 |
| 7 | 0.0954 | 47.71 | 64.30 | 0.0468 | 0.6422 | 1.416 | 0.829 | 0.198 | 11.272 |

r) Carriers data set.( http://lib.stat.cmu.edu/datasets/).

The data arose in a study to develop screening methods to identify carriers of a rare genetic disorder. It consists of two groups. Four measurement were made on blood samples. The age of patients was also included. The most appropriate number of clusters appears to be $k=2$ suggested by CH, Silhouette coefficient, $D_1$, $\tau_2$, but the classification of the entities is not the expected one. The criterion $Min\{|\boldsymbol{W}(\gamma)\}$ separated out seven patients which are more a chance aggregation of points in the PC's space than a real pattern in the data.



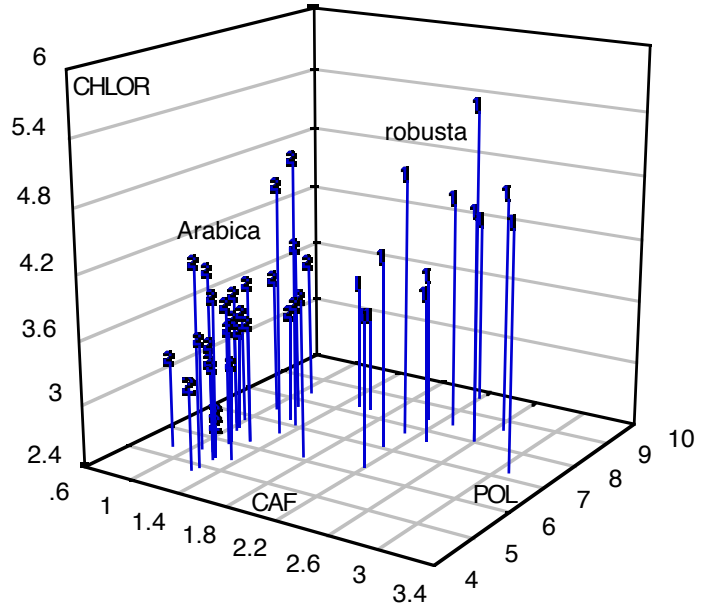

In general, the two-cluster case may be difficult to detect because many data set have at least one reasonable split even if no clusters is in effect present. In this particular case the task is above the possibilities of **DetClus** which, although has been successfully applied in numerous empirical studies fails under certain conditions. In particular, when the cluster have a very different shape the results of **DetClus** can be poor or meaningless.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 26.903 | 32.9 | 223 | 1.076 | 0.897 | 0.46 | 0.78 | 0.14 | 9.85 |
| 3 | 7.627 | 34.9 | 218 | 0.686 | 0.635 | 1.01 | 0.94 | 0.10 | 10.35 |
| 4 | 3.514 | 34.4 | 143 | 0.562 | 0.458 | 1.32 | 0.92 | 0.07 | 10.65 |
| 5 | 1.791 | 29.1 | 245 | 0.448 | 0.524 | 0.93 | 0.84 | 0.17 | 10.63 |
| 6 | 0.907 | 26.2 | 250 | 0.327 | 0.457 | 1.51 | 0.87 | 0.19 | 10.71 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 7 | 187 | | | | |
| 3 | 143 | 6 | 45 | | | |
| 4 | 58 | 6 | 31 | 99 | | |
| 5 | 114 | 2 | 4 | 13 | 61 | |
| 6 | 88 | 2 | 5 | 12 | 29 | 58 |

s) Coffee Data set

The two most important varieties of commercial coffee are *Coffea arabica* and *Coffea canephora*, usually known as arabica and robusta, respectively. Commercial coffee beverage is made from arabica or robusta beans or blends of them, the arabica being considered of better quality and is therefore more expensive. Martin *et al*. (1998) determined the content of chlorogenic acid, caffeine, trigonelline, amino acids, polyphenols and aqueous extract have been in *n=41* samples of green coffee (13 robust and 28 arabic). Arabica and robusta varieties from different geographic origin were included. The following figure shows the entities in the plane formed by caffeine, chlorogenetic acid and total polyphenol.




| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | 28 | 13 | | | | |
| 3 | 19 | 11 | 1 | | | |
| 4 | 6 | 20 | 7 | 8 | | |
| 5 | 6 | 11 | 8 | 9 | 7 | |
| 6 | 12 | 6 | 4 | 8 | 3 | 8 |

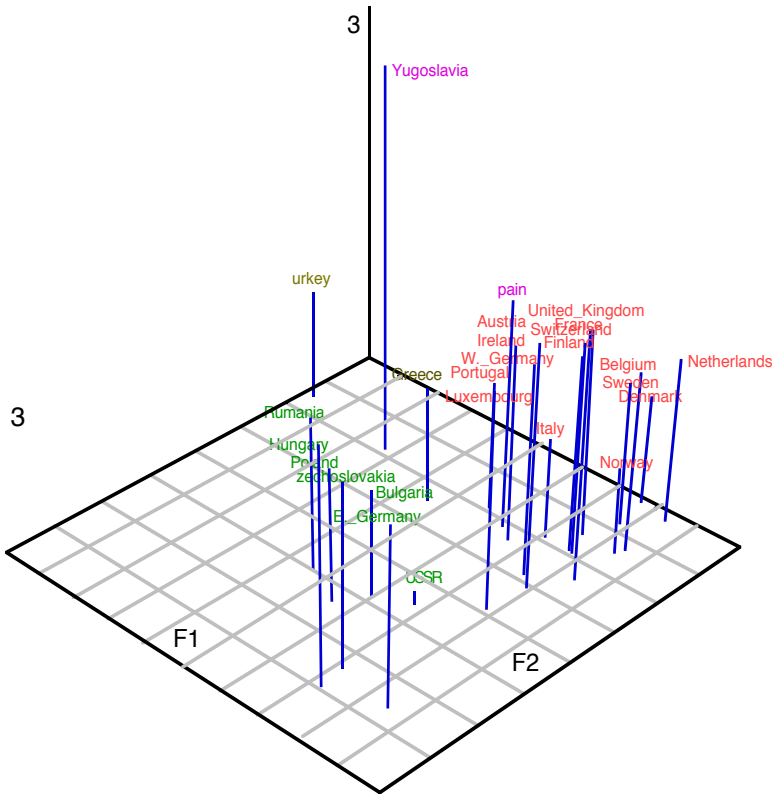
| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 14.064 | 15.7 | 36.4 | 0.563 | 0.802 | 3.24 | 0.68 | 0.287 | 7.60 |
| 3 | 5.687 | 17.3 | 42.2 | 0.512 | 0.641 | 2.14 | 0.85 | 0.177 | 8.26 |
| 4 | 2.277 | 10.2 | 29.6 | 0.364 | 0.716 | 1.90 | 0.63 | 0.247 | 8.71 |
| 5 | 1.334 | 9.5 | 25.6 | 0.333 | 0.635 | 2.30 | 0.49 | 0.238 | 9.01 |
| 6 | 0.698 | 7.3 | 20.9 | 0.251 | 0.700 | 2.05 | 0.33 | 0.344 | 9.25 |

For *k=2* the recovery rate is 100%. A two-cluster solution is proposed by the silhouette coefficient, $D_1$, and $\tau_2$. The appropriate number of clusters is *k=3* for the Calinski index, CH, and $D_1$. The choice *k=4* is advocated by Marriott. This choice is also a second-best for the silhouette coefficient, $D_2$ and $\tau_2$. It appears that there is a subdivision of each species into two subclusters, but their relevance is an open question.

t) European jobs data file ( http://lib.stat.cmu.edu/datasets/).

The data are the percentage employed in different industries in Europe countries during 1979. The job categories are agriculture, mining, manufacturing, power supply industries, construction, service industries, finance, social and personal services, transport and communications.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 15 | 11 | | | | | | |
| 3 | 7 | 2 | 17 | | | | | |
| 4 | 2 | 12 | 7 | 5 | | | | |
| 5 | 2 | 1 | 4 | 12 | 7 | | | |
| 6 | 3 | 1 | 12 | 4 | 2 | 4 | | |
| 7 | 1 | 1 | 4 | 12 | 3 | 1 | 4 | |
| 8 | 1 | 1 | 4 | 4 | 9 | 3 | 1 | 3 |

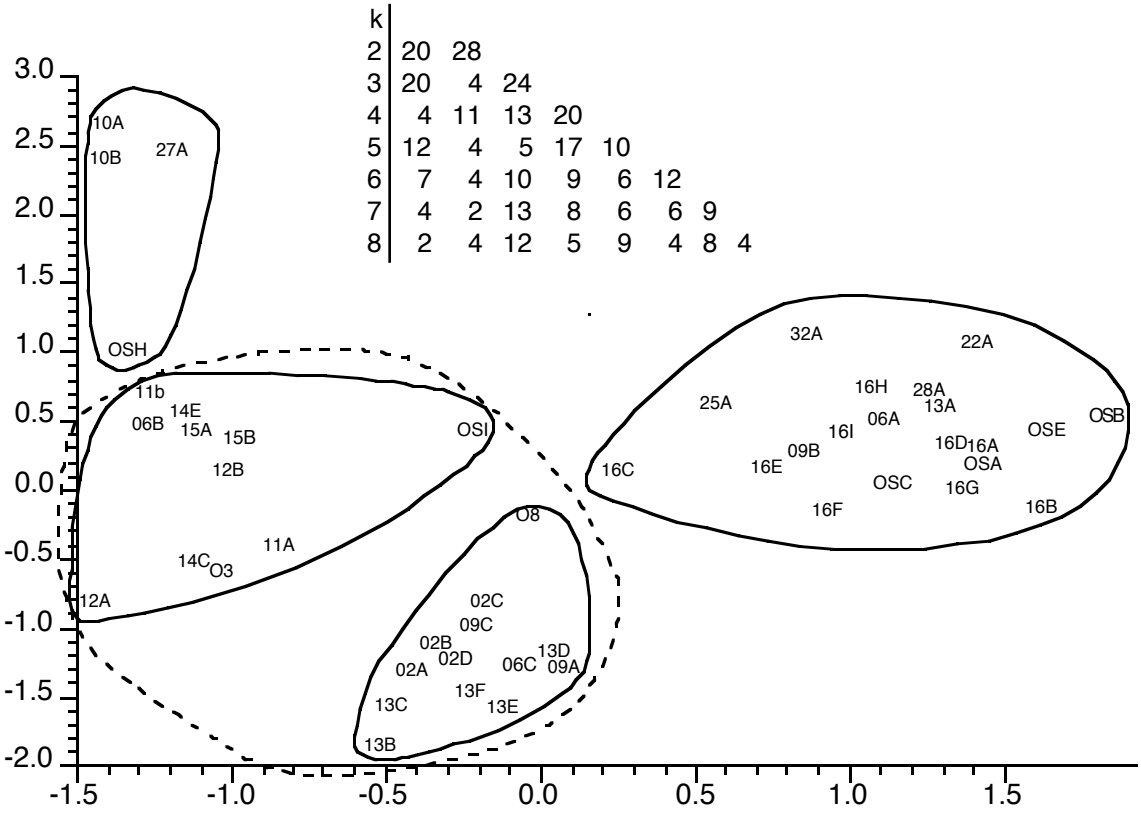



The analysis shows that the countries cluster together into two main groups along political lines: Group 1 contains the countries of the communist East Bloc (these data were collected during the Cold War). Group 2 contains countries of capitalist Western Europe. Moreover, there are two small clusters: (Spain,Yugoslavia) wich shared some characteristics of both groups, and (Greece,Turkey), which were not aligned with the European standards. This classification is confirmed by the Calinski-CH index which shows two contiguous peaks at *k=4* and *k=5*. The silhouette coefficient indicates *k=5*.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 9.6128 | 4.26 | 4.26 | 0.3845 | 0.751 | 8.079 | 0.799 | 0.904 | 8.609 |
| 3 | 1.4472 | 4.74 | 4.74 | 0.1302 | 0.800 | 4.573 | 0.622 | 1.705 | 8.480 |
| 4 | 0.2096 | 5.15 | 5.15 | 0.0335 | 0.786 | 4.524 | 0.659 | 1.902 | 8.552 |
| 5 | 0.0516 | 5.16 | 5.16 | 0.0129 | 0.833 | 7.386 | 0.506 | 2.816 | 8.708 |
| 6 | 0.0144 | 4.95 | 4.95 | 0.0052 | 0.826 | 3.904 | 0.500 | 3.553 | 8.880 |
| 7 | 0.0033 | 4.55 | 4.55 | 0.0016 | 0.858 | 3.096 | 0.371 | 7.020 | 9.053 |
| 8 | 0.0008 | 4.39 | 4.39 | 0.0005 | 0.807 | 2.151 | 0.435 | 10.656 | 9.186 |

u) Roman pottery data set

Roman Terra sigillata was produced in central Italy since the first century B.C. and then traded and produced throughout the Roman world. The composition of *n=48* shreds of terra sigillata was analyzed in Pop *et al*. (1995) by using seven elements: K, Mg, Ca, Ti, Mn, Fe, Al). The data are represented in the space of the first principal components explaining 74% of total variation.




**DetClus** was run on the first *m=4* principal components. The four-cluster solution is indicated by Calinski-CH, and (at least partly) by the Marriot measure. The Dunn index and the Davies-Bouldin index advocate *k=5*. The silhouette coefficient and $C^3$ have a critical point at *k=3*. These results are most likely due the fact that the classes do not have sharp boundaries. Moreover, a good partition (according to Pop *et al*, 1995) for this data set, has a very large (at least 13) number of clusters ad doubtful attributions are expected.

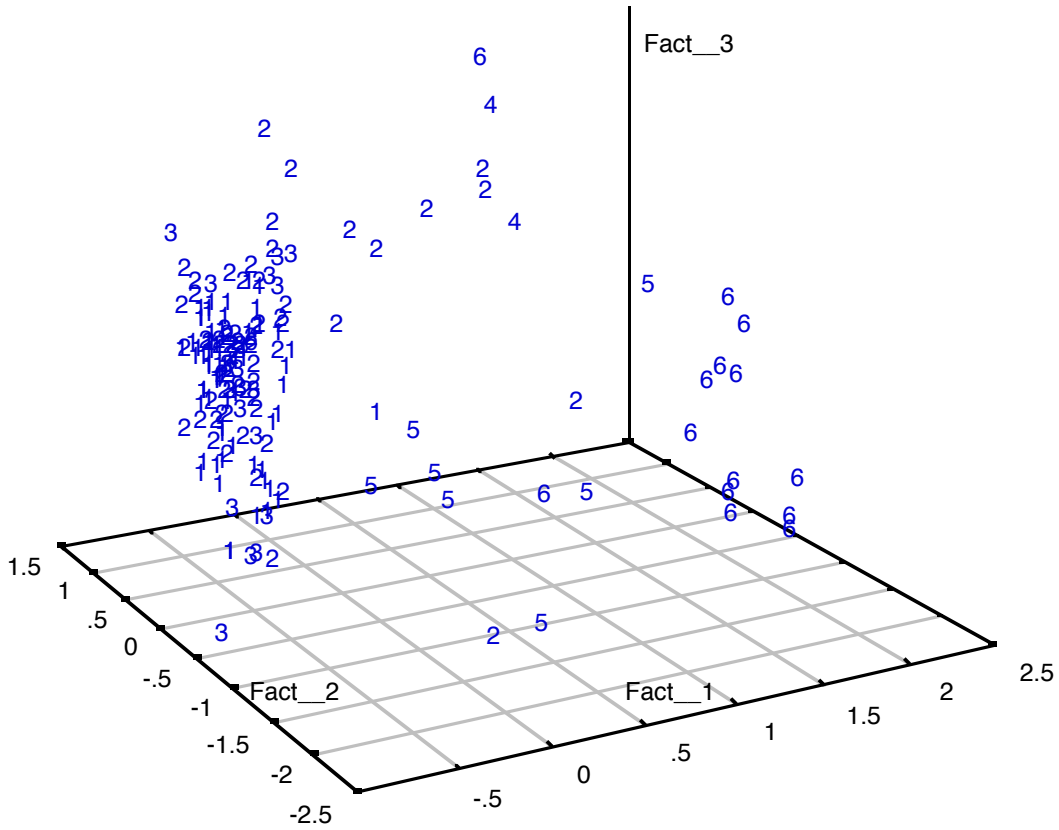
| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 8.0443 | 13.73 | 13.73 | 0.3218 | 0.837 | 30.731 | 0.559 | 0.415 | 9.019 |
| 3 | 0.7983 | 18.57 | 18.57 | 0.0718 | 0.845 | 12.183 | 0.505 | 0.918 | 8.878 |
| 4 | 0.1518 | 27.60 | 27.60 | 0.0243 | 0.823 | 11.047 | 0.557 | 0.883 | 8.901 |
| 5 | 0.0614 | 22.21 | 22.21 | 0.0153 | 0.801 | 5.470 | 0.653 | 0.771 | 9.168 |
| 6 | 0.0285 | 27.17 | 27.17 | 0.0103 | 0.728 | 7.352 | 0.782 | 0.699 | 9.277 |
| 7 | 0.0135 | 24.68 | 24.68 | 0.0066 | 0.756 | 4.541 | 0.674 | 0.910 | 9.441 |
| 8 | 0.0076 | 23.88 | 23.88 | 0.0049 | 0.735 | 3.112 | 0.722 | 0.790 | 9.564 |

v) Glass identification database (Blake and Merz, 1998)

The study of classification of type of glass was motivated by criminological investigation. The data set has 214 entities and 9 attributes measured on a ratio scale: refractive index, sodium, magnesium, aluminum, silicon ,potassium, calcium, barium, iron. There are two bigger clusters (163 window glass and 51 non-window glass) and both the clusters can be separated in 3 subclusters: WG=70 building windows, 17 vehicle windows, 76 building windows; NWG=13 containers, 9 tableware, 29 headlamps.

**DetClu**s has been applied to the first four principal components after eliminating variable 1 and 9. The recovery rate is generally less than satisfactory. No stopping rule was able to detect the true number of cluster *k=6* (a possible exception are Marriott and $D_1$)

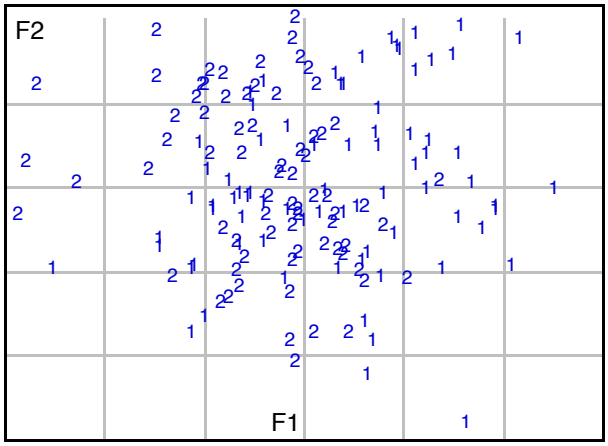
| k | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 50 | 164 | | | | | | | |
| 3 | 157 | 26 | 31 | | | | | | |
| 4 | 161 | 31 | 2 | 20 | | | | | |
| 5 | 127 | 38 | 16 | 2 | 31 | | | | |
| 6 | 23 | 18 | 18 | 2 | 31 | 122 | | | |
| 7 | 41 | 9 | 2 | 110 | 11 | 18 | 23 | | |
| 8 | 3 | 11 | 23 | 2 | 33 | 18 | 115 | 9 | |
| 9 | 16 | 3 | 9 | 114 | 7 | 29 | 23 | 11 | 2 |
| 10 | 23 | 34 | 13 | 11 | 24 | 3 | 2 | 77 | 9 | 18 |

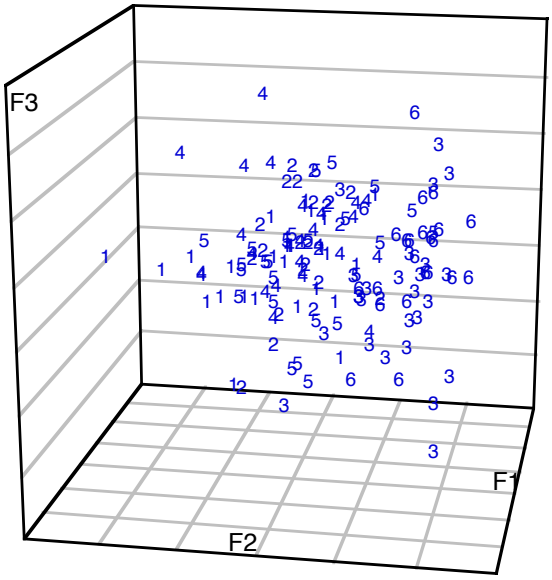



| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 11.509 | 60.2 | 60.2 | 0.460 | 0.836 | 2.86 | 0.68 | 0.065 | 11.26 |
| 3 | 3.537 | 48.4 | 48.4 | 0.318 | 0.816 | 1.64 | 0.74 | 0.060 | 11.29 |
| 4 | 1.505 | 72.9 | 72.9 | 0.241 | 0.829 | 0.46 | 0.53 | 0.212 | 11.05 |
| 5 | 0.653 | 83.3 | 83.3 | 0.163 | 0.667 | 0.97 | 0.43 | 0.243 | 11.03 |
| 6 | 0.324 | 65.6 | 65.6 | 0.117 | 0.644 | 0.79 | 0.61 | 0.185 | 11.16 |
| 7 | 0.178 | 61.6 | 61.6 | 0.087 | 0.621 | 1.06 | 0.34 | 0.203 | 11.20 |
| 8 | 0.106 | 67.9 | 67.9 | 0.068 | 0.667 | 0.42 | 0.53 | 0.199 | 11.18 |
| 9 | 0.071 | 68.0 | 68.0 | 0.057 | 0.690 | 0.50 | 0.11 | 0.213 | 11.20 |
| 10 | 0.051 | 59.5 | 59.5 | 0.051 | 0.563 | 1.00 | 0.58 | 0.217 | 11.24 |

w) Grey kangaroos data set (Andrews and Herzberg, 1985).

The data analyzed are *m= 18* skull measurements on *n= 148* reference animals of known sex and species : *Macropus fuliginous giganteus* (25 females, 25 males), *M.f. melanops* (25f, 23m), *M.f. fuliginous* (25f, 25m). Three variables: palate width, mandible length, super-occipital- parocciptal depth were discarded because of the excessive number of missing values. The other missing values were estimated as the group mean. The principal component analysis were applied to reduce the number of variables retaining 9 factors explaining 98.1% of total variation. The presence of a cluster structure is doubtful. The classification by sex appears reasonable, but the six classes (sex by species) were not recognized by **DetClus**.



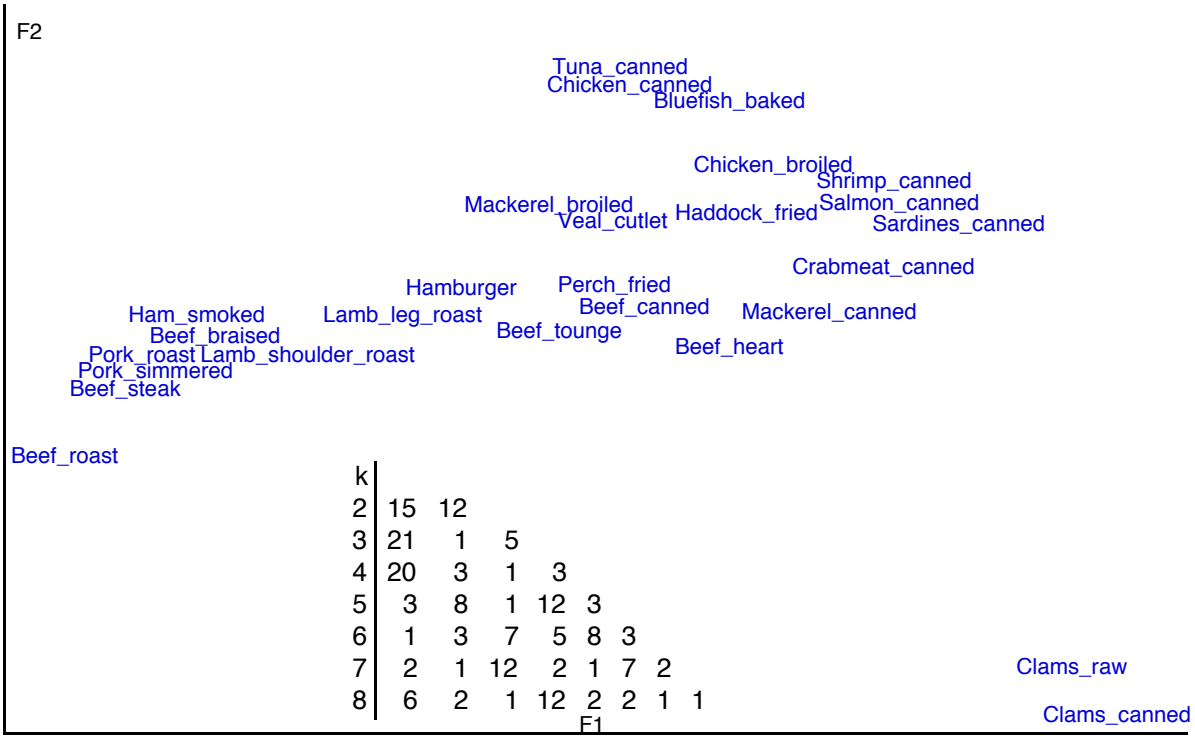




| k | Criterion | Calinski | CH | C-index | Marriott | Si.Co. | D1 | D2 | D3 | τ1 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 21.442 | 13.96 | 13.96 | 0.1844 | 0.858 | 0.448 | 1.08 | 1.51 | 0.513 | 0.76 | 0.057 | 11.972 |
| 3 | 7.781 | 11.31 | 11.31 | 0.1067 | 0.700 | 0.375 | 1.24 | 1.17 | 0.704 | 0.73 | 0.047 | 11.903 |
| 4 | 3.748 | 10.79 | 10.79 | 0.0843 | 0.600 | 0.343 | 1.07 | 1.37 | 0.745 | 0.56 | 0.051 | 11.877 |
| 5 | 1.956 | 10.84 | 10.84 | 0.0701 | 0.489 | 0.349 | 0.69 | 1.45 | 0.774 | 0.47 | 0.055 | 11.861 |
| 6 | 1.095 | 10.29 | 10.29 | 0.0568 | 0.394 | 0.369 | 0.79 | 1.19 | 0.810 | 0.30 | 0.061 | 11.901 |
| 7 | 0.695 | 10.16 | 10.16 | 0.0657 | 0.340 | 0.327 | 1.47 | 1.37 | 0.771 | 0.33 | 0.057 | 11.914 |
| 8 | 0.468 | 9.39 | 9.39 | 0.0519 | 0.299 | 0.364 | 0.54 | 1.08 | 0.798 | 0.48 | 0.078 | 11.968 |
| 9 | 0.290 | 9.56 | 9.56 | 0.0455 | 0.235 | 0.357 | 0.68 | 1.17 | 0.825 | 0.35 | 0.072 | 11.954 |
| 10 | 0.223 | 8.84 | 8.84 | 0.0507 | 0.223 | 0.356 | 0.59 | 1.17 | 0.789 | 0.40 | 0.090 | 11.998 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 94 | 54 | | | | | | | | |
| 3 | 59 | 47 | 42 | | | | | | | |
| 4 | 36 | 58 | 17 | 37 | | | | | | |
| 5 | 18 | 17 | 31 | 36 | 46 | | | | | |
| 6 | 27 | 11 | 29 | 35 | 13 | 33 | | | | |
| 7 | 21 | 27 | 21 | 14 | 18 | 23 | 24 | | | |
| 8 | 27 | 10 | 20 | 4 | 27 | 30 | 16 | 14 | | |
| 9 | 25 | 14 | 25 | 10 | 12 | 9 | 22 | 22 | 9 | |
| 10 | 3 | 7 | 12 | 8 | 15 | 24 | 28 | 17 | 25 | 9 |

x) %Nutrient data from yearbook of Agriculture 1959, USA Department of Agriculture, Washington, DC (Hartigan, 1974, p.86)

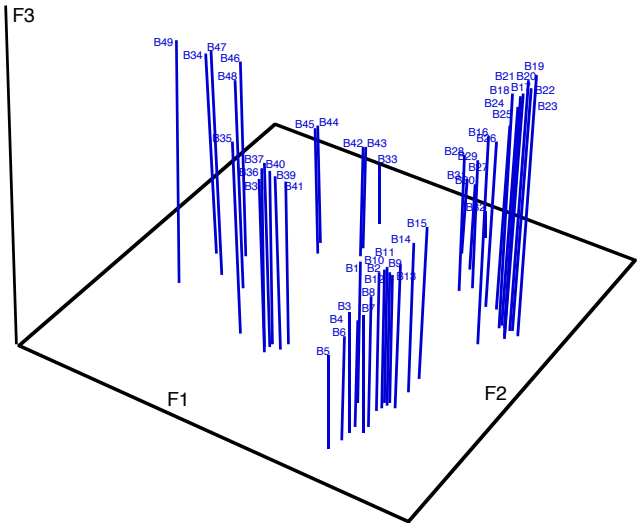
This data depicts nutrients of different kinds of meat, fish and fowl. Five features, namely food-energy, protein, fat, calcium and iron were used for Clustering. **Detclus** was run on the first four principal components explaining 99.97% of total variation. The figure below reports the factor scores of the two dominant components (66.84% of variance explained).




According to Hartigan, plausible stopping points in the clustering are *k=2*, *k=5*, and *k=8*. Only the Marriott statistic states clearly that the underlying number of cluster is *k=5* (however, the $C^3$ obtains the smallest increment just for *k=5*). The Calinski-CH and $D_2$, indicate k=6 (but *k=5* shows an interesting subpeak). The statistic $\tau_2$ and $D_1$ suggest *k=4* (and perhaps the Marriott). The silhouette coefficient indicates *k=3*. Only an expert of the data can find the most appropriate choice in the interval 3-6.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 13.9951 | 6.85 | 6.85 | 0.5598 | 0.739 | 6.340 | 0.802 | 0.437 | 8.175 |
| 3 | 1.7348 | 6.11 | 6.11 | 0.1561 | 0.907 | 6.527 | 0.387 | 4.444 | 8.296 |
| 4 | 0.2446 | 7.06 | 7.06 | 0.0391 | 0.892 | 3.291 | 0.406 | 6.477 | 8.370 |
| 5 | 0.0313 | 10.22 | 10.22 | 0.0078 | 0.859 | 11.400 | 0.418 | 5.892 | 8.399 |
| 6 | 0.0100 | 11.05 | 11.05 | 0.0036 | 0.809 | 10.104 | 0.501 | 4.899 | 8.559 |
| 7 | 0.0021 | 9.00 | 9.00 | 0.0010 | 0.770 | 1.925 | 0.366 | 38.416 | 8.773 |
| 8 | 0.0005 | 10.8 | 10.8 | 0.0003 | 0.877 | 2.611 | 0.302 | 36.560 | 8.866 |

y) Abernethy Forest 1974 data (Gordon, 1999, p.46). As an aid to identifying past vegetation n=49 cores of sediment were taken from Abernethy Forest in northeast Scotland. The pollen spectrum is described by the percentage of the *m=9* most important pollen types. ***DetClus*** was run on the first five principal components (representing 92.7% of the variability contained in the data set). According to Gordon (1999, p. 45) there are *k=5* clusters.




How to decide on the optimal number of clusters? Calinski-CH confirm *k=5* which is the best guess also for $D_1$, $D_2$, $\tau_2$, and silhouette coefficient (this has a valley in the graph for *k=5*). The $C^3$ criterion has a major increment after *k=5* which can be interpreted as a sign of a good value for *k*.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2.02671 | 11.45 | 11.45 | 0.08107 | 0.963 | 19.33 | 0.29 | 3.96 | 9.30 |
| 3 | 0.13553 | 14.19 | 14.19 | 0.01220 | 0.962 | 7.69 | 0.33 | 3.45 | 9.15 |
| 4 | 0.00873 | 19.55 | 19.55 | 0.00140 | 0.947 | 8.25 | 0.37 | 3.06 | 9.15 |
| 5 | 0.00067 | 32.86 | 32.86 | 0.00017 | 0.916 | 15.04 | 0.39 | 2.47 | 9.17 |
| 6 | 0.00014 | 30.50 | 30.50 | 0.00005 | 0.919 | 9.53 | 0.36 | 2.80 | 9.37 |
| 7 | 0.00003 | 77.49 | 77.49 | 0.00002 | 0.888 | 12.54 | 0.43 | 2.48 | 9.41 |
| 8 | 0.00001 | 75.4 | 75.4 | 0.00001 | 0.859 | 10.67 | 0.48 | 2.17 | 9.56 |
| 9 | 0.00000 | 60.8 | 60.8 | 0.00000 | 0.860 | 12.10 | 0.52 | 1.56 | 9.69 |

| k | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 37 | 12 | | | | | | |
| 3 | 7 | 37 | 5 | | | | | |
| 4 | 5 | 7 | 32 | 5 | | | | |
| 5 | 7 | 17 | 5 | 5 | 15 | | | |
| 6 | 5 | 7 | 15 | 3 | 17 | 2 | | |
| 7 | 11 | 5 | 6 | 15 | 3 | 7 | 2 | |
| 8 | 7 | 3 | 11 | 6 | 6 | 2 | 9 | 5 |
| 9 | 2 | 15 | 2 | 11 | 5 | 2 | 3 | 3 | 6 |

z) Economics of Cities

Prices and earnings around the globe Economic Research Department, Union Bank of Switzerland, Zurich. The data represent the economic conditions in 48 cities around in world in 1991. There are m=3 variables: weighted average of the number of working hours in 12 occupations, index of the cost 112 goods and services excluding rent (Zurich = 100), index of hourly earnings in 12 occupations after deductions (Zurich = 100).


Apparently, there are six groups. This is suggested by the Calinski, CH and $D_2$. The Marriott index and the silhouette coefficient for this data set are hermetic. $D_1$ indicates $k=4$ or $k=7$; the separation index $\tau_2$ advocates $k=5$.

| k | Criterion | Calinski | CH | Marriott | Si.Co. | D1 | D2 | τ2 | C3 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 19.9310 | 16.02 | 22.79 | 0.797 | 0.701 | 5.99 | 0.856 | 0.137 | 8.46 |
| 3 | 7.7809 | 18.82 | 8.23 | 0.700 | 0.665 | 2.89 | 0.883 | 0.143 | 9.33 |
| 4 | 3.5537 | 16.53 | 21.26 | 0.569 | 0.652 | 2.46 | 0.821 | 0.149 | 8.95 |
| 5 | 1.4096 | 19.72 | 24.11 | 0.352 | 0.684 | 9.16 | 0.755 | 0.211 | 9.07 |
| 6 | 0.8421 | 22.25 | 37.15 | 0.303 | 0.646 | 7.02 | 0.859 | 0.179 | 9.11 |
| 7 | 0.5063 | 20.37 | 33.65 | 0.248 | 0.672 | 2.79 | 0.833 | 0.194 | 9.60 |
| 8 | 0.3168 | 17.92 | 24.65 | 0.203 | 0.686 | 5.74 | 0.746 | 0.224 | 9.76 |
| 9 | 0.1751 | 12.51 | 13.74 | 0.142 | 0.763 | 2.63 | 0.577 | 0.419 | 9.96 |
| 10 | 0.0999 | 15.62 | 22.04 | 0.100 | 0.746 | 3.57 | 0.632 | 0.455 | 9.92 |

| k | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 25 | 21 | | | | | | | |
| 3 | 20 | 6 | 20 | | | | | | |
| 4 | 16 | 5 | 6 | 19 | | | | | |
| 5 | 5 | 11 | 9 | 6 | 15 | | | | |
| 6 | 5 | 5 | 11 | 7 | 9 | 9 | | | |
| 7 | 6 | 5 | 9 | 7 | 11 | 5 | 3 | | |
| 8 | 4 | 3 | 7 | 7 | 7 | 6 | 6 | 6 | |
| 9 | 5 | 5 | 6 | 8 | 2 | # | 1 | 5 | 3 |
| 10 | 1 | 5 | 8 | 6 | 7 | 5 | 4 | 2 | 5 | 3 |

### 4.2.4 Summary

The findings in this limited set of examples lack probative value in showing the conduct of **DetClus** for a specific application, but shed light on the tendency to cluster of the entities that are subject of the clustering in question. The results can be summarized as follows.

1) It is advisable to plot the value of the indices against the number of clusters, and assessing the plot by eye, looking for peaks and valleys and consensus among the various indices.

2) The values of $k$ to be investigated should be kept small as some indices tend to assume a bizarre behavior as $k \rightarrow n$.

3) The user should consider both peaks and valleys as an indication of a good candidate values for $k$.

4) If the clustering is strong enough to withstand true confusion and plots used with discretion then there is a good chance of determining the true value of k by considering the point on which all indices are concordant.

5) In general, when partial agreement is found between stopping rules, then the user should opt for the smaller number of clusters.

For a fixed quality of the initial partition the recovery rate tends to be better for even sized cluster than for cluster of unequal size

***DetClus*** is an iterative partitioning method (or k-means) which aims to trace closed non intersecting boundaries surrounding high density zones separated from other such zones by zones containing low density of points. With strongly clustered data, almost any clustering approach yields significant insights and **DetClus** has given satisfactory practical results although it is possible that for a particular application it will perform poorly. In fact, a good starting partition, an optimal reassigning scheme, faithfully reproduction of the algorithm in the programming language, and a high speed computer do not guarantee that the end partition will coincide with one of the best configuration compatible with the data set. On the other hand, if this is the case we have no way of knowing it, unless the structure of the data were perfectly known as in the case of artificial data sets used in simulation.

**DetClus** is not practical when it comes to solving large problems as all iterative schemes take a lot of time to get a solution and require a large amount of space for storage purposes. It take about two weeks to execute the largest problem *(n=12000, m=50, k=2,3,...,25).*

## 5. Syntax

Although the literature on cluster analysis insistently advises the neophyte to avoid writing a computer program on his/her own (Cormack, 1971; Gnanadesikan et al. 1977; Zupan, 1981; Blashfield *et al.* 1982, Digby and Kempton, 1987) I could not resist the pleasure of doing a little programming by myself. Firstly, because most of the steps in the algorithm compelled to combine and harmonize a number of subroutines usually run separately. Secondly, because the maximal dimensions allowed by standard packages were largely lower than what was needed for a complete analysis of several data sets. Thirdly, because most of the homemade statistical software is for the PC's keeping the Macinthosh out of the realm of the "number crunching" (Le Progiciel R 4.0 by P. Casgrain-P. Legendre, 2001 and the Vista by F. Young,1999 provide noteworthy exceptions). Finally, because the type of recursive computation I had in mind sees more visibility in text books and journal articles than in readily available computer programs constraining the data analyst to concentrate on the mechanics of implementation instead of the result interpretation.

The above considerations have led to the development of the ***DetClus*** software described in this manual.

The **DetClus** procedure performs a disjoint cluster analysis on the basis of Mahalanobis distances computed from a fixed set of metric variables. The algorithm start with an arbitrary choice of a feasible classification of the entities into clusters (this means that every entity belongs to one and only one cluster) Then, keeping the same number of clusters, a sequence of possible reassignments is considered. The reassignment that yields the maximum benefit is made and the process is repeated until an optimum of the criterion is reached.

The program is written in FutureBasic3. It can run on PPC under system classic. Several new ideas are included in the present manual: for example, that of a relocations of the entities based on a global best improving strategy, a preliminary estimation of the within-cluster dispersion matrix enhancing the performance of a large variety of initialization methods, a system of induces for the choice of the number of clusters, and so on. The program has been successfully applied in a number of data sets, although there is a need for more extensive study to compare the opportunities offered by ***DetClus*** with other commercial and freeware software and better understand its strengths and limitations. In view of the important statistical role of the k-means, it is hoped that the software presented herein will be useful in many applications.

No publication can cover all requirements for all users of cluster analysis. ***DetClus*** seeks to provide users with the material necessary to perform an iterative partition (Friedman-Rubin approach) on their own data. The present version, freely available on the internet network, can handle a maximum of $n=12000$ data points, $m=50$ variables, and $k=25$ clusters. Other versions can be requested if the program does not handle enough data for your applications (and you have enough memory storage and execution time). *Please send bug reports to agotar@unical.it*

The variables are not transformed so that any transformation of the raw data should be executed before using **DetClus**.

Missing values are not allowed. The user is expected eliminate all cases with one or more missing values.

***DetClus*** satisfies the constrain 2b so that each entity is admitted to the cluster it has its smallest distance with, even if this distance is large. As a consequence, the software does not indicate which entities are obviously members of a cluster and which should be regarded as borderline entities, singletons or outliers. The user, however, is free to remove the suspected entities from the data set. *The default values are* $k_1=2$ *and* $k_2=Min\{[(n-m(m-1))/(3m)]+1,25\}$

**The format of the data set**

***DetClus*** does not accept mixed data types so that binary, nominal and ordinal variables having numeric values will be considered as metric

The program can only process a rectangular matrix *(n x m)* entities-by-variables matrix and a dissimilarity matrix will not be accepted as input file.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{21} \\ \vdots & \vdots & \ddots & & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \ddots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix}$$

There are three file specification to be given

1. The input data file (according to the scheme *rows=entities, columns=variables*;
2. The output file (contains the results of the run);
3. The clusterings file (contains the cluster membership of all the entities for each value of *k*).

Additional specifications are required if the user provides a set of initial centroids or a starting configuration or a set of cardinalities. A separate file is also necessary for the true or known classification (for a fixed number of clusters) if comparison must be performed with the final partition of the algorithm.

***DetClus*** requires a separate line (or set of lines) for each entity, containing the entity identification and variable values, utilizing a comma delimited format

$$(AAA\ldots AAA, Value\_1,\ Value\_2,\ \ldots,\ Value\_m)_i, \qquad i = 1,2,\ldots,n$$

The field *"AAA...AAA"* expects a maximum of 32 letters or numbers or special characters, but considers the symbols to form a name and not a value on which regular computations can be performed. A labels can have less than 32 characters (at least one is required). The comma cannot be used as part of a case name. The data Value_1, Value_2 etc. must be numeric. The data in the file should be formatted as text items delimited by commas and/or carriage-returns. Each item is assigned to a separate variable.

**DetClus** reads a line of text from the file, beginning at the current "file mark" position (which is usually at the beginning of the line), and ending when a carriage-return character is encountered, or the end of the file is encountered, or 255 characters have been read, whichever occurs first. The file mark is then advanced to a position just past the last character read. **DetClus** then attempts to assign each of the comma-delimited items in the text line to one of the variables (Value_1, Value_2 etc.) in the variable list. If there are more items in the text line than variables in the list, the extra items are discarded. If there are fewer items in the text line than variables in the list, then zeros are assigned to the extra variables.

The variables should be in free-field format with valued separated by a comma. A comma is also required between the case name and the value of the first variable. The same position in successive data rows need not contain value for the same variable. Only the order of variable specification must be identical.

```
Setosa,51.000,35.000,14.000,2.000
Setosa,49.000,30.000,14.000,2.000
----------------------------------------
```

**DetClus** cannot read Excel files. In Excel save the data in a text file in tab separated format (or space separated or separated any other delimiter) and then, if necessary, open the file with a text editor and change the delimiter to a comma.

All numeric data are considered in double precision. It is important that the data row does not end with a comma and that a comma is not included in the entity label because the pairing value of the variable/datum on the row will be shifted. The consequence of this error may be dramatic. Sometimes an input impairment occurs allowing the detection of the input errors; otherwise, particularly if the identification labels are numbers, the run will be executed and blindly accepted if the findings look like what expected. The file which contains the initial partition and that of the true classification file (the two specifications may coincide) have only one integer column (the decimal point, if present, will be ignored).

Unfortunately, the user interface of **DetClus** is the part of the program that has received scant attention thus far. The Lions's share of effort were received by the initialization and relocation methods.

The procedure should be considered as a heuristic approach and clusters determined at its end partition have to be considered with caution until there is enough evidence of their existence *e.g*. external information justifying the tacit assumption of underlying normal distributions with equal dispersion matrices containing approximately equal proportion of entities. It must be emphasized that **DetClus,** as the other clustering procedures, may impose a grouping structure on the data set *D* even though *D* may not possess such a structure. In the latter case, the clusters found are not a genuine content of the data, but a synthetic product of the procedure which can constitute a very misleading description of the data set. *In other words, cluster analysis is not a panacea. That is, we must have an indication that the vector of D form clusters before we apply a clustering algorithm* (Theodoridis and Koutroumbas, 1998, p. 543).

While reasonable effort has been made to ensure this software operates substantially as described the Author cannot guarantee proper operation in every possible configuration. For this reason, the **DetClus** is provided "As is" without warranty of any kind, either expressed or impled, including, but not limited to, the implied warranties of merchantability, fitness for a particular purpose, or non-infringement. In no event will Agostino Tarsitano be liable for any special, incidental, consequential, indirect or similar damages due to loss of data or any other reason.

If a user encounters problems with this program, the author is willing to help solve them.

.

**References**


Al-Daoud M.B. Roberts S.A. (1996). New methods for the initialization of clusters. *Pattern Recognition Letters*, 17, 451-455.

Art D. , Gnanadesikan R., Kettenring J.R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21A, 75-99.

Anderson E.J. (1996). Mechanisms for local search. E*uropean Journal of Operational Research*, 88, 139-151.

Andrews D.F. Herzberg A.M. (1985). Data. A collection of problems from many fields for the student and researh worker. Springer-Verlag, New York. (http://lib.stat.cmu.edu/DASL/)

Arnold S.J. (1979). A test for clusters. *Journal of Marketing Research*, 16, 545-551.

Banfield J.D. Raftery A.E.(1993). Model-based Gaussian and non Gaussian clustering. *Biometrika*, 49, 803-828

Bayley T.A., Dubes R. (1982. Cluster validity profiles. *Pattern Recognition*, 15, 61-83.

Bayne C.K., Beauchamp J.J., Begovich C.L., Kane V.E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition* 12, 51-62.

Beale E.M.L. (1972). Euclidean Cluster Analysis. *Bulletin of the International Statistical Institute*, 43, 92-94.

Bezdek J.C. Pal N.R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B,* 28,301-315.

Bissell A.F.(1976). Ordered random selction without replacement. *Applied Statistics*, 35,73-75.

Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. http://www.ics.uci.edu/~mlearn/MLRepository.html

Blashfield R.K., Aldenderfer M.S., Morey L.C. (1982). Cluster Analysis Software. P.R. Krishnaiah and L.N. Kanal eds. Handbook of Statistics. Vol.2, 245-266. North-Holland Publishing Company.

Bock H.H. (1998). Probability models and hypothesuis testing in partitioning cluster analysis. In P.Arabie, J. Hubert & G. De Soete (Eds.), Clustering and Classification (pp. 377-453). World Scientific Publishing, Singapore.

Calinski T. (1969). On the application of cluster analysis to experimental results. *Bulletin of the International Statistical Insitute,* 43, 101-103.

Casgrain P. (2001). Le Progiciel R v4.0d6. http://ProgicielR.webhop.org/

Castagnoli E. (1978). Un'osservazione sull'analisi classificatoria. Seminario su due temi di analisi statistica multivariata. CLEUP, Bologna (I), 135-139.

Cormack R.M. (1971) *A Review of Classification*. Journal of the Royal Statistical Society. Series A, Vol. 134, 321-353.

Chen H., Gnanadesikan R., Kettenring J.R. (1974). Statistical methods for grouping corporations. *Sankhyã*, 36, series B, 1-28.

Chernoff H. (1970). Metric considerations in cluster analysis. Pp. 621-629 in: *Proceedings of the 6h Berkley symposium on mathematical statistics and probability*, UCLA Press, Berkley, CA.

Chernoff H. (1973) The use of faces to represent points in k-dimenional space graphically. *Journal of the American Statistical Association*, 73, 361-368.

Coleman D., Dong X. ,Hardin J. ,Rocke D.M. ,Woodruff D.L. (1999). Some computational issue in cluster analysis with no a priori metric. *Computational Statistics & Data Analysis*, 31, 1-11.

Cox D.R. (1957). Note on grouping. *Journal of the American Statistical Association*, 52, 543-547.

Dagnelie P. Merckx A. (1991). Using generalized distances in classification of groups. *Biometrical Journal*, 33, 683-695.

Davenport M. Studdert-Kennedy G. (1972). The Statistical Analysis of Aesthetic Judgements: An Exploration. *Applied Statistics*, 21, 324-333.

Davies D.L. Bouldin D.W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224-227.

Digby P.G.N. Kempton R.A. (1987. Multivariate Analysis of Ecological Communities. Chapman & Hall, London.

Dubes R.C. (1987): How many cluters are best? An experiment. *Pattern Recognitio*n, 20, 645-663.

Duda R.O. Hart P.E. (1973). Pattern classification and scene analysis. John Wiley & Sons, New York.

Duda R.O. Hart P.E. Stork D.G. (2001). Pattern classification. 2nd editions. John Wiley & Sons, New York.

Engelman L. Hartigan J.A. (1959). Percentage points of a test for clusters.*journal of the American Statistical Association*, 64, 1647-1648.

Everitt B.S. (1979) Unresolved probles in cluster analysis. *Biometrics,*35, 169-181.

Everitt B.S., Landau S., Lase M. (2001). Cluster analysis, 4th ed. Arnold, London.

Fisher L. Van Ness J.W. (1971) Amissible Clustering Procedures. *Biometrika,* 58, 91-104.

Fisher D., Xu L., Zard N. (1992). Ordering effects in clustering. In: *Proceedings of the9th International Conference on MachineLearning*, pp. 163-168. Morgan Kauffman, San Matero, Ca.

Forgy E. (1965). Cluster analysis of multivariate data. Efficiency vs interpretability of classification. WNAR Meetings, UCLS Riverside, Ca. June 22-23, 1965 (Abstract in *Biometrics*, 21, 768).

Friedman H.P. Rubin J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

Gentleman J.F. (1975). Algorithm AS 88. Generation of all $_{n}C_{r}$ combinations by symulating nested Fortran Do loops. *Applied Statistics*, 24, 374-376 (http://www.Stat.unipg.it/pub/stat/statlib/apstat)

Gnanadesikan R., Kettenring J.R., Landwehr J.M. (1977). Interpreting and assessing the results of cluster analysis. *Bulletin of the International Statistical Institute*, 47, 451-463.

Gordon A.D. (1996). How many clusters? An investigation of five procedures for detecting nested cluster structure. In C. Hayashi et al. (Eds.). *Data science, classification, and related methods*. Proceeedings of the fifth conference of the international federation of classification societies, pp 109-116. Kobe, Japan, March, 27-30, 1996.

Gordon A.D. (1999). Classification, 2nd Edition. Chapman and Hall, London.

Hall D.J. Khanna D. (1977). The ISODATA method computation for the relative perception of similarities and differences in complex and real data. In K. Enslein, A. Ralston, and H.W. Wilf (eds.) *Statistical Method for Digital Computers*. John Wiley & Sons. New York.

Hand D.J.,Daly F., Lunn A.D., McConway K.J., OstrowskiE. (Editors)(1994). Handbook of small data sets. Chapman & Hall, New York.

http://www2.ncsu.edu/eos/service/pams/stat/sicl/small_data/index.html

Hansen P. Mladenovic (2001). J-means: a non local search heuristic for minimum sum of squares clustering. *Pattern recognition,* 34, 405-413.

Hardy A.(1996). On the number of clusters. *Computational Statistics and Data Analysis*, 23, 83-96.

Hartigan J.A.Wong M.A. (1979). A k-Means clustering algorithm. *Applied Statistics*,28, 100-108.

Heeler R.M. Day G.S. (1975). A supplementary note on the use of cluster analysis for stratification. *Applied Statistics*, 3, 342-343.

Hubert L.J., Arabie P. (1985). Comparing partitions. *Journal of Classification*, 2, 193-218.

Hubert L.J Levin j.R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83, 233-253

Ismail M.A., Kamel M.S. (1989). Multidimensional data classification utilizing hybrid search strategies. *Pattern Recognition*, 22, 75-89.

Kaufman L. Rouseeauw P.L. (1990). Finding Group in Data. An Introduction to Cluster Analysis. John Wiley & Sons, New York.

Kennard R.W., Stone L.A. (1969). Computer aided design of experiments. *Technometrics*. 11, 137-148.

Klastorin T.D. (1983). Assessing cluster analysis results. *Journal of Marketing Research*, 20, 92-98.

Knuth D.E. (1981). The Art of Computer Programming. Vol.2, 2nd Edition. Addison-Wesley Publishing Company, Reading.

Kotari R; Pitts D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20, 405-416.

Kruskal J. (1977). Multidimensional scalind and clustering. In Classification and Clustering. J. Van Ryzin (Ed.). Academic Press, New York. 17-44.

Lance G. Williams W.T. (1967*).* A general theory of classificatory aorting strategies. II. Clustering systems. *Computer Journal,* 10, 271-277.

Lubishew A.A. (1962): On the Use of Discriminant Function in Taxonomy. *Biometrics,* 18, 455-477..

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. In:L.M. LeCam, J. Neyman (Eds.) *Proceedings of the 5th Berkeley sysmposium on mathematical statistics and probability, Vol. 1-Statistics* pp. 281-297, UCLA Press, Berkeley, Ca.

Mahalanobis P.C. (1936). On the generalizedistance in statistics. *Proceedings of the National Institute of Science* (India), 12, 49-55.

Manly B.F.J. (1985). The statistics of natural selection on animal populations. Chapman and 5Hall, London.

Maronna R. Jacovkis (1974). Multivariate Clustering Procedure with Variable Metrics. *Biometrics,* 30, 499-505.

Marriott F.H.C. (1971). Practical Problems in a Method of Cluster Analysis. *Biometrics*, 27, 501-514.

Martin M.J. Pablos F. Gonzalesz A.G. (1998). Discrimination between arabica and robusta green coffee varieties according to their chemical composition. *Talanta*, 46, 1259-1264.

McRae D.J. (1971). MIKCA: A Fortran IV Iterative k-means Cluster Analysis Program. Behavioral Science. Vol. 16, 423-424.

Meza J.C. Olkin I. (1993). Numerical procedures for estimanting the parameters in a multivariate homogeneous correlation model with unequal variances. *Sankhya, A,* 506-515.

Milligan G.W. (1980). An examination of the effect of six types of error perturbation on fifteen 5clustering algorithms. *Psychometrika,* 45, 325-342.

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

Milligan G.W.(1996). Clustering Validation: Results and Implications for Applied Analysis. In P.Arabie, J. Hubert & G. De Soete (Eds.), Clustering and Classification (pp. 341-373). World Scientific Publishing, Singapore.

Mineo A. (1985) A New Criterion for the Choice of Seed Points for a Nearest Centroid Cluster Analysis. *Rivista di Statistica Applicata*, 18,191-198.

O'Flaherty and MacKenzie (1982). Algorithm As172: direct simulation of nested fortran Do-Loops. Applied Statistics, 31, 71-74.

Peña J.M., Lozano J.A., Larrañaga P. (1999). An empirical comparison of four initialization

methods for the k-means algorithm. *Pattern Recognition Letters*, 20, 1027-1040.

Pop F. Dumitrescu D.Sarbu C. (1995). A study of roman pottery (terra sigillata) using hieararchical fuzzy clustering. *Analytica Chimica Acta*, 310, 269-279.

Rencher A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. The American Statisticiam, 217-225.

Rao M.M. (1971). Cluster Analysis and Mathematical Programming. *Journal of the American Statistical Association*, 66, 622-626.

Rasson J.P. Kubushishi T. (1996). The gap test: an optimal method for determining the number of natural classes in cluster analysis.

Richards L.E. (1972). Refinement and Extensions of Distribution-free Discrimant Analysis. *Applied Statistcs*, 21, 174-176.

Rouncefield M. (1995). The statistics on poverty and inequality. *Journal of Statistics Education*, 3. http://lib.stat.cmu.edu/DASL/DataArchive.html

Rousseeuw P.J. Kaufman L. Trauwaert E. (1996). Fuzzy clustering using scatter matrices. Computational statistics and data analysis. 23, 135-151.

Rubin J. (1967). Optimal classification into groups: an approach for solving the taxonomy problem. *Journal of Theoretical Biology*, 15, 103-144.

Sadocchi S. (1977). Un metodo di cluster analysis non stratificato derivato dal metodo del legame singolo. *Rivista di Statistica Applicata*,10, 228-241.

Sarle W.S. (1983). Cubic clustering criterion. SAS technical Report A+108, Cary, Nc: SAS institute Inc.
http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/hrddoc/indfiles/5903.htm

Scott A.J., Symons M.J. (1971). Clustering method based on likelihood ratio criteria. *Biometrics*, 27, 387-397.

Symons M.J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*. Vol. 37, 35-43.

Seber G.A.F. (1984). Multivariate Observations. John Wiley & Sons. New York.

Selim S.Z., Ismail M.A. (1984). K-means type algorithm: generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 81-87.

Späth H. (1985). Cluster Dissection and Analysis. Theory, FORTRAN Programming, Examples. Ellis Horward Limited, Chichester (U.K.).

Struyf A. Hubert M., Rousseeuw P.J. (1997).Integrating robust clustering techniques in S-Plus. *Computational Statistics & Data Analysis*, 26, 17-37

Tarsitano A. Anania G. (1995) Tecniche di analisi statistica multivariata per l'individuazione dei sistemi agricoli territoriali in Italia. G. Cannata (a cura di). I sistemi territorali agricoli italiani degli anni '90. Contributi metodologici. Rubettino Editore, Soveria Mannelli (Cz), Italy.

Tarsitano A. (2002). A computational study of several relocation methods for k-means algorithms. Submitted for publication.

Theodoris S. Koutroumbas K. (1998). Pattern recognition. Academic Press

Van Rijsbergen (1970). A clustering algorithm. *Computer Journal,* 13, 113-115.

Vicari D. (1990). Indici per la scelta del numero dei gruppi (indices for determining the numebr of clusters). *Metron*,47, 473-492.

Vichi M. (1985). Cluster analysis and the graphical approach for identification of two types of multivariate analysis. *Metron*, 43, 165-188.

Warkenar C.S. Krishna G. (1979): A heuristic clustering algorithm using union of overlapping pattern-cells. *Pattern Recognition* 11, 85-93.

Wong C. Che C. Su M. (2001). A novel algorithm for data clustering. *Pattern recognition*, 34, 425-442.

Zhang Q., Boyle R.D. (1991). A new clustering algorithm with multiple runs of iterative procedures. *Pattern Recognition*, 24, 835-848.

Zupan J. (1981). Clustering large data sets. John Wiley & Sons, New York.