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).

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

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,…,k,; i≠j*. When data set is well-structured the proper number of cluster determines a maximum of *S(i)* near the upper bound *(S(i)≤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

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

where the sums are over *r,s= 1,2,…,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

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

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

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

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

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

ρ_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.

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

η_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.

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

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

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

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.

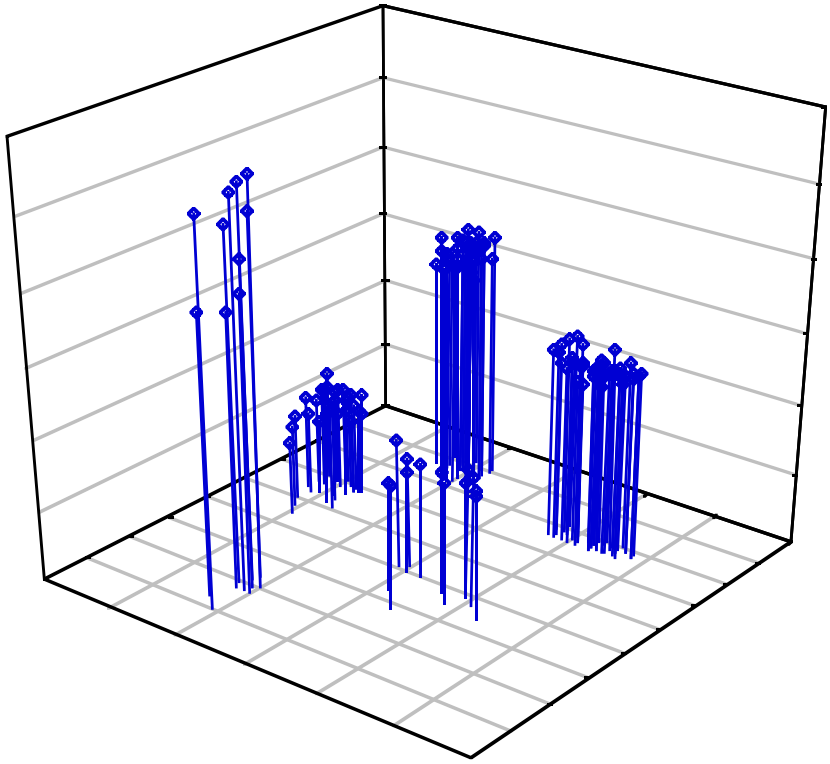
poorly separated clustered (i-z)

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.



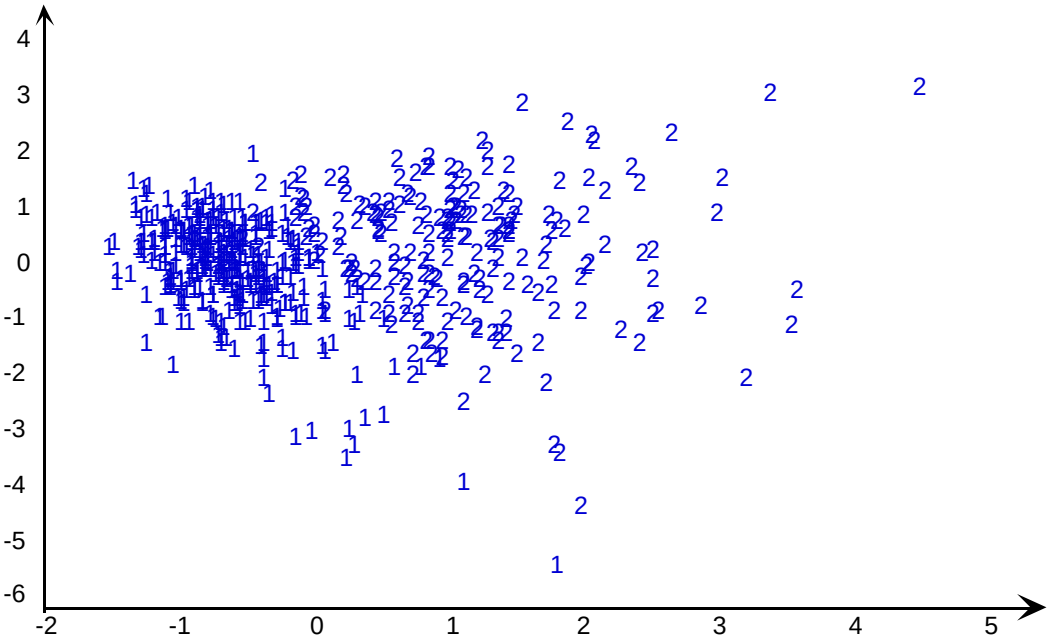
k	Criterion	Calinski	CH	Marriott	Si.Co.	D1	τ2	C3
2	1.10521	34.0	125.6	0.04421	0.9748	99.81	2.20	9.70
3	0.10151	41.6	131.5	0.00914	0.9612	6.68	1.36	10.07
4	0.01100	56.3	115.6	0.00176	0.9533	3.73	1.15	10.32
5	0.00125	48.9	297.2	0.00031	0.9257	15.00	0.90	10.33
6	0.00013	85.5	729.9	0.00005	0.9340	8.84	1.04	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.



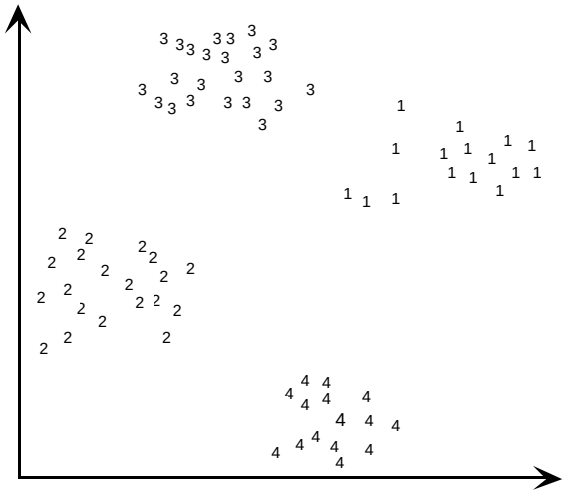
For *k=2* the recovery rate is 100%: $R_{HA}=1$ and *Q=1*. The correct number of cluster is indicated by all the indices

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 τ_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 μ_i and μ_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

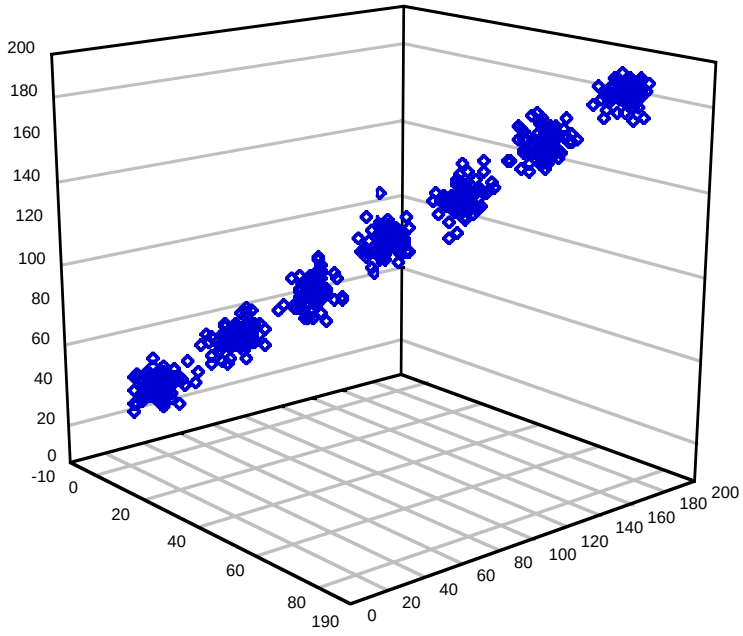
All the applicable indices have a crucial point (a local minimum or a local maximum) at *k=5*.

.

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



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

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

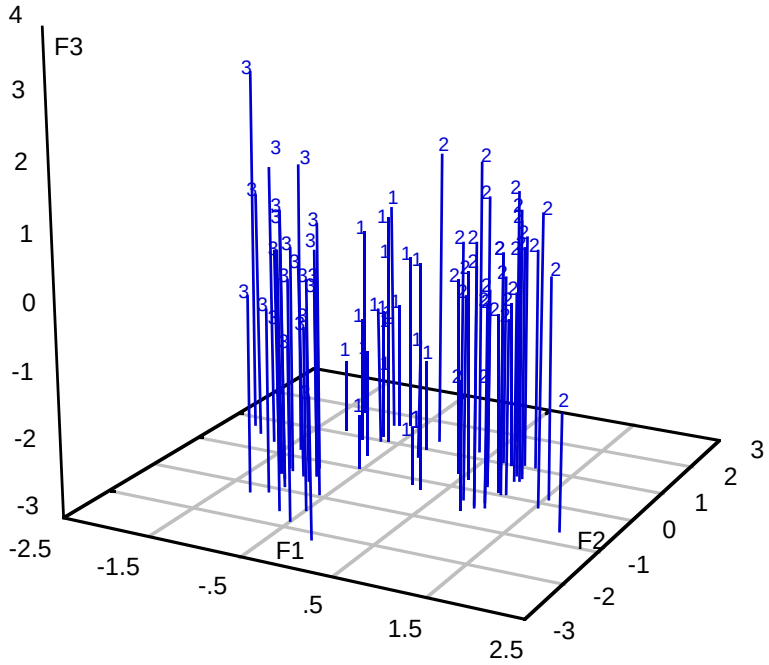
The intracluster covariance matrix has homogeneous correlations and distinct variances. This structure has been applied to model phenomenon that are permutation invariant, as for example, in the case of *m* equivalent psychological tests (Meza and Olkin, 1993). Since the data set is strongly clustered,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



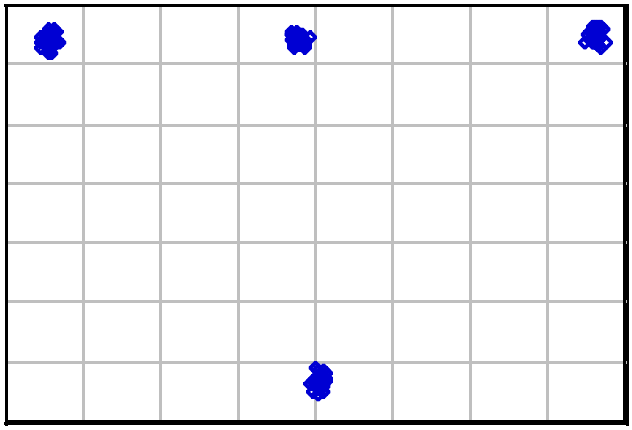
The data set has contains three natural clusters which can be clearly recovered. The value *k=3* is indicated by CH (which coincides with the Calinski), Marriott, C-index, D_3, and τ_2. The S. coefficient, D_1, and τ_1 indicate *k=2* or *k=3*. The index C^3 a minimum at *k=3*. It appears that both a 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.



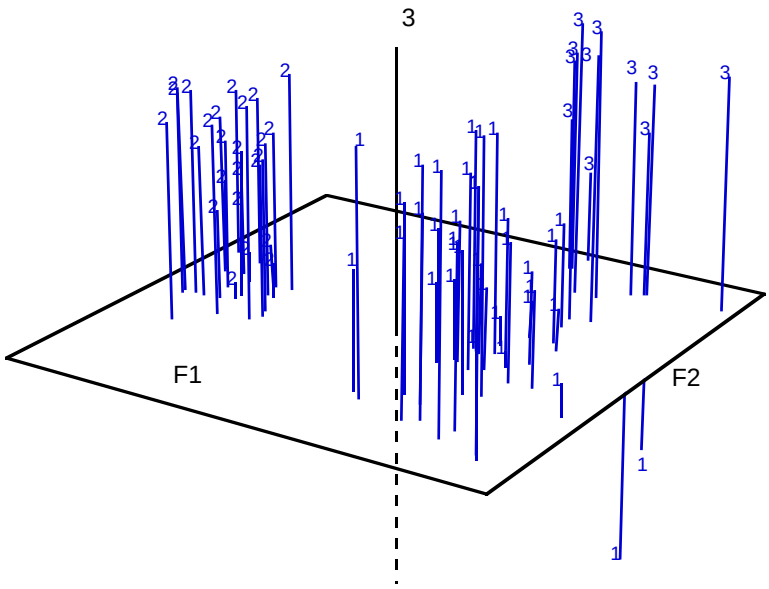
The true value of *k* has been detected by almost all the indices. Only C^3 gives a wrong indication. The two-cluster solution is considered a reasonable value for the number of clusters by the C-index, Silhouette coefficient, D_1 and τ_2. This is not surprising if one considers the structure of the 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 τ_1 gives *k=2* although there is an interesting strong decrease after *k=3*.

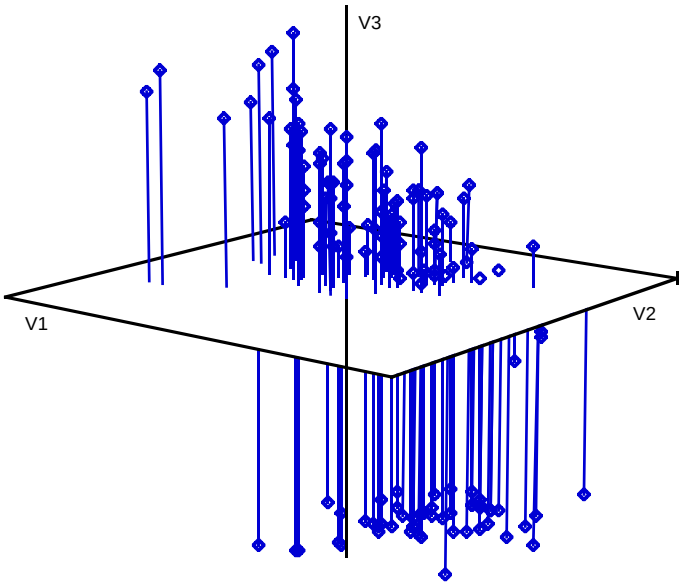
k	Criterion	Calinski	CH	Marriott	Si.Co.	D1	τ2	C3
2	7.40866	25.6	25.6	0.29635	0.8603	3.98	0.31	9.93
3	1.31860	32.4	32.4	0.11867	0.8285	1.63	0.28	9.78
4	0.53943	30.9	30.9	0.08631	0.7199	2.26	0.22	9.93
5	0.31099	36.8	36.8	0.07775	0.7165	0.84	0.20	9.99
6	0.17675	30.5	30.5	0.06363	0.5778	1.10	0.12	10.16
7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



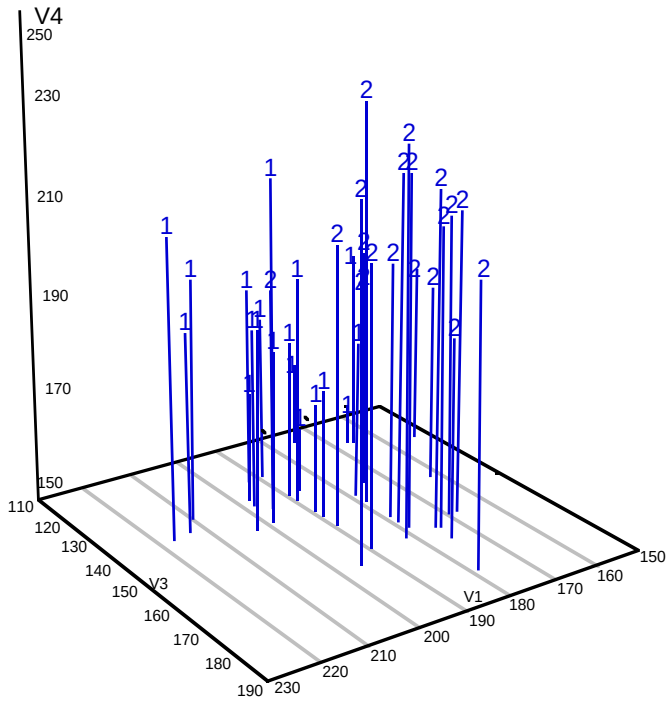
k							
2	100	50					
3	50	49	51				
4	34	18	48	50			
5	50	12	29	27	32		
6	50	28	19	23	18	12	
7	18	25	12	21	26	23	25

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

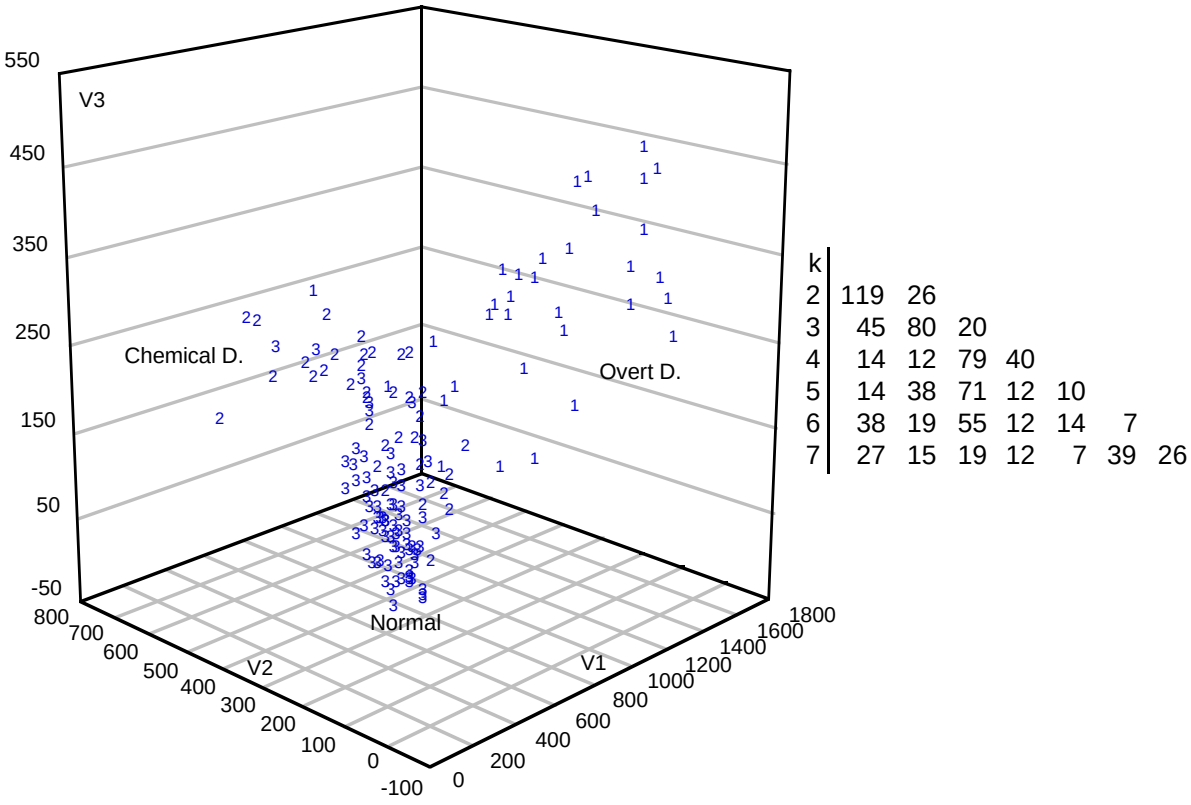
In this case the dispersion matrices are very different for the two types and it is doubtful whether it is even legitimate to suppose they have a common dispersion matrix. Nevertheless, **DetClus** retrieved almost exactly the two genies except for 2 *H. oleracea* and 1 *H. cardiourum*. The recovery 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 τ_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.



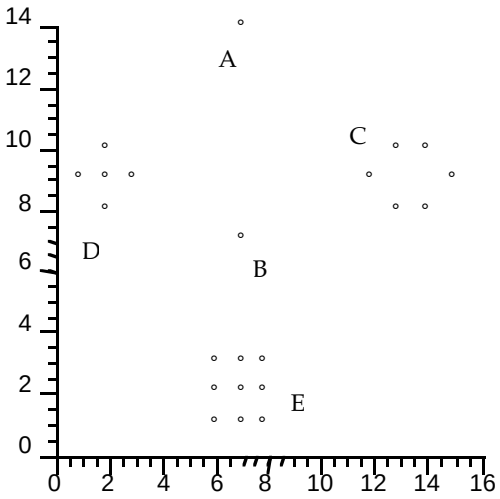
k							
2	119	26					
3	45	80	20				
4	14	12	79	40			
5	14	38	71	12	10		
6	38	19	55	12	14	7	
7	27	15	19	12	7	39	26

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 τ_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, τ_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.

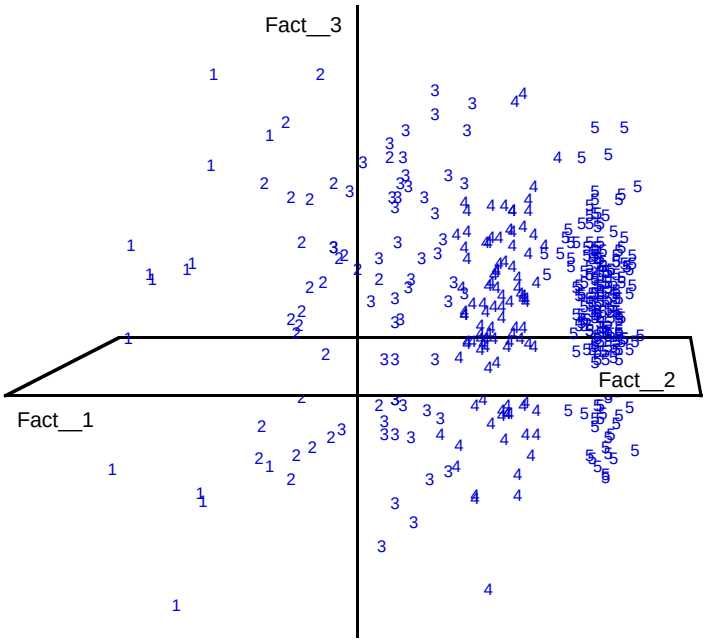
m) Clusters of unequal size and dispersion

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 τ_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.

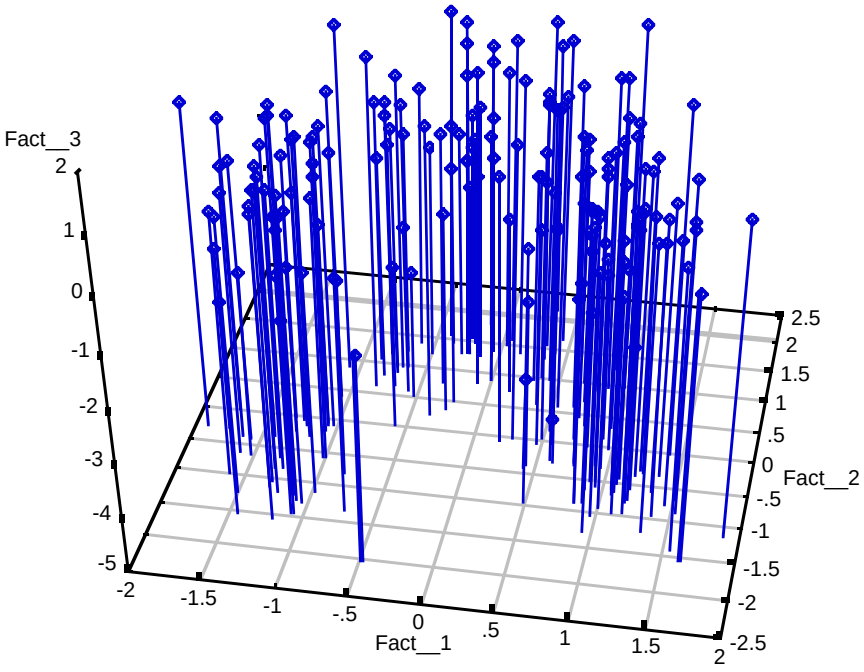


k								
2	346	119						
3	152	63	250					
4	240	19	137	69				
5	55	240	16	124	30			
6	113	16	65	107	128	36		
7	114	17	47	125	35	56	71	
8	15	144	30	29	96	26	67	58
9	56	25	80	74	16	86	35	30	63

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.

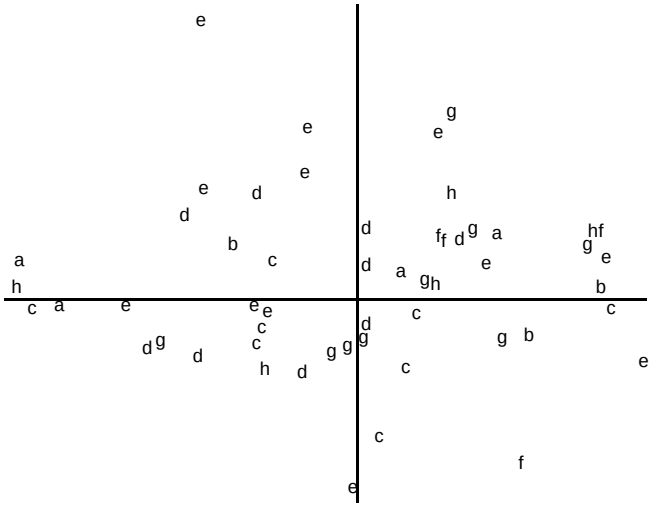


The true number of clusters was found by Calinski, CH, and Marriott. A local peak is found at *k=2* for the silhouette coefficient and τ_2. The Dunn index and the Davies-Bouldin yield *k=5*. It must be noted that the final solution for *k=3* has a modified Rand index of $R_{HA}=0.9188$ with 4 entities 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 τ_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.).

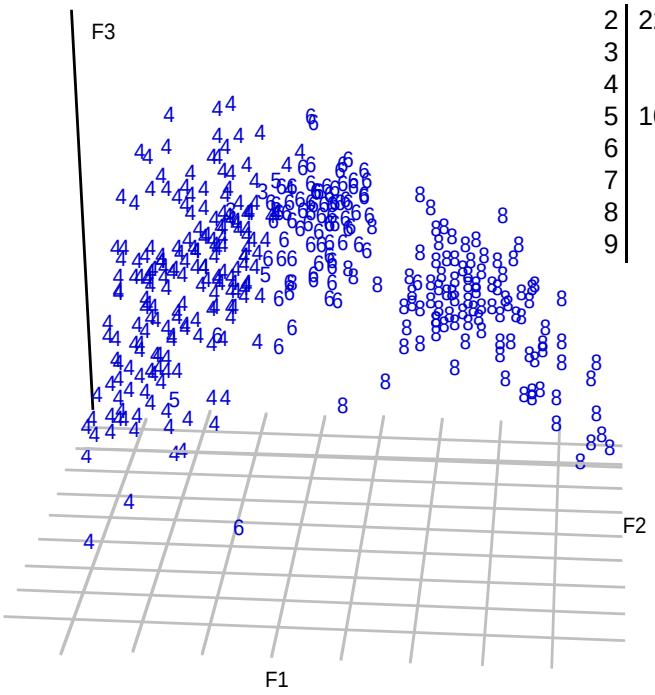
2	22	32							
3	21	21	12						
4	11	10	19	14					
5	8	20	13	10	3				
6	11	9	2	9	11	12			
7	2	7	10	6	6	11	12		
8	5	11	5	2	10	8	9	4	
9	5	7	6	6	3	12	5	7	3
10	6	5	2	5	4	4	11	5	4	8

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.



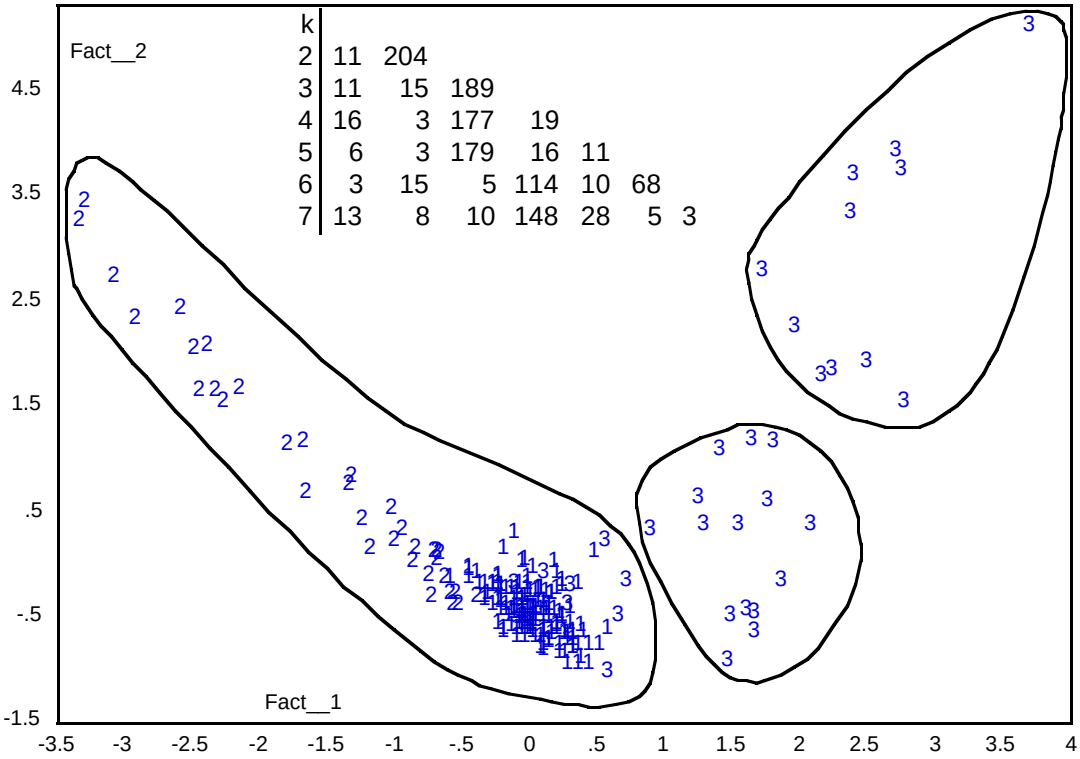
2	225	167							
3	73	98	221						
4	72	125	98	97					
5	100	76	23	72	121				
6	22	98	71	76	53	72			
7	32	22	76	53	72	86	51		
8	82	71	17	5	33	77	50	57	
9	85	21	59	15	53	51	71	5	32

The indication on the underlying number of clusters is discordant across the indices: Calinski, silhouette coefficient, τ_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.

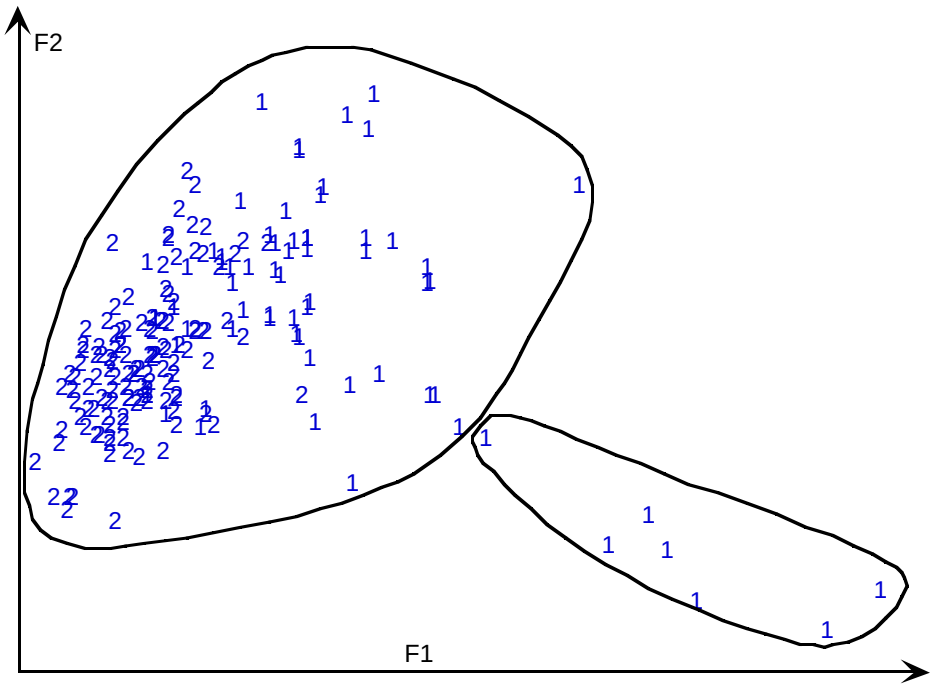


Nevertheless, the value *k=3* sounds reasonable for the Calinski index, CH, D_2 and C^3. Marriott, silhouette coefficient D_1 and τ_2 suggest *k=5* which does not appear consistent with these data. It is worth of note the fact that the true number of clusters can be successfully detected even if the 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, τ_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.



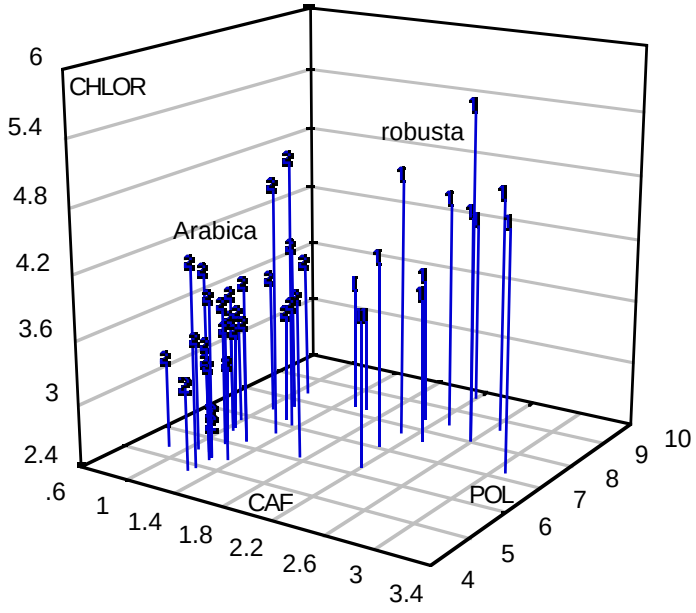
In general, the two-cluster case may be difficult to detect because many data set have at least one reasonable split even if no clusters is in effect present. In this particular case the task is above the possibilities of **DetClus** which, although has been successfully applied in numerous empirical 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.



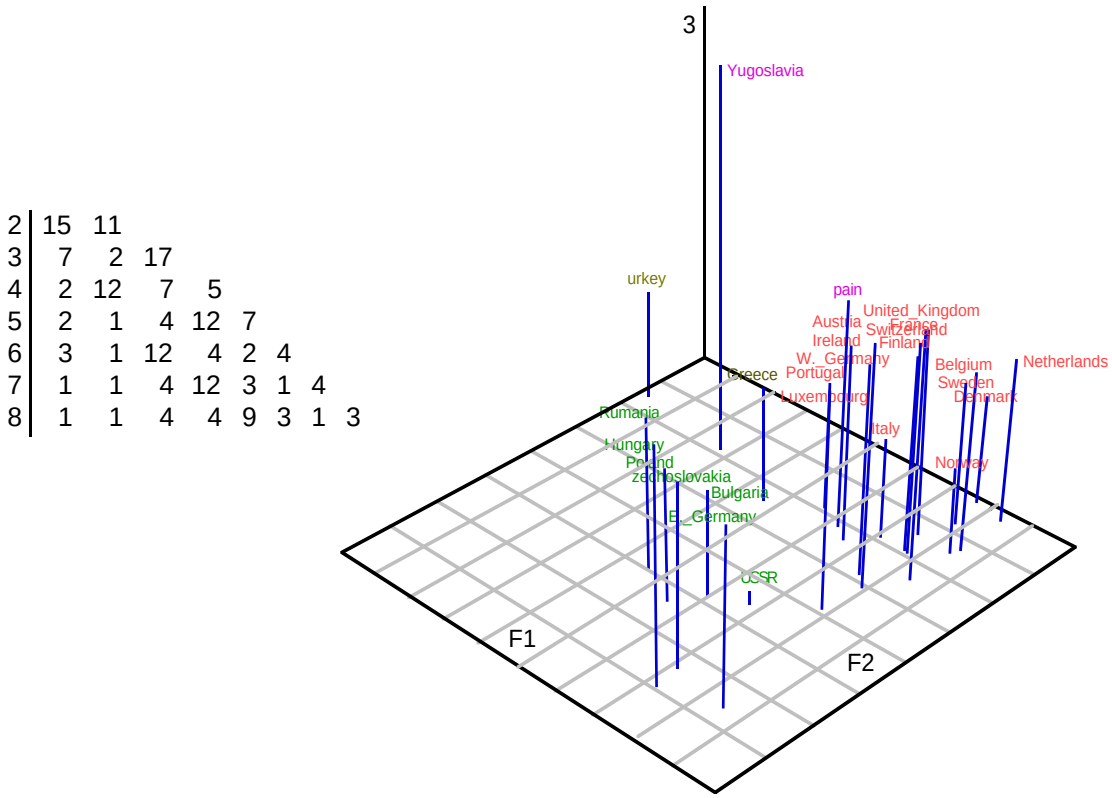

```
2 | 28  13
3 | 19  11   1
4 |  6  20   7  8
5 |  6  11   8  9  7
6 | 12   6   4  8  3  8
```

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 τ_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 τ_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.

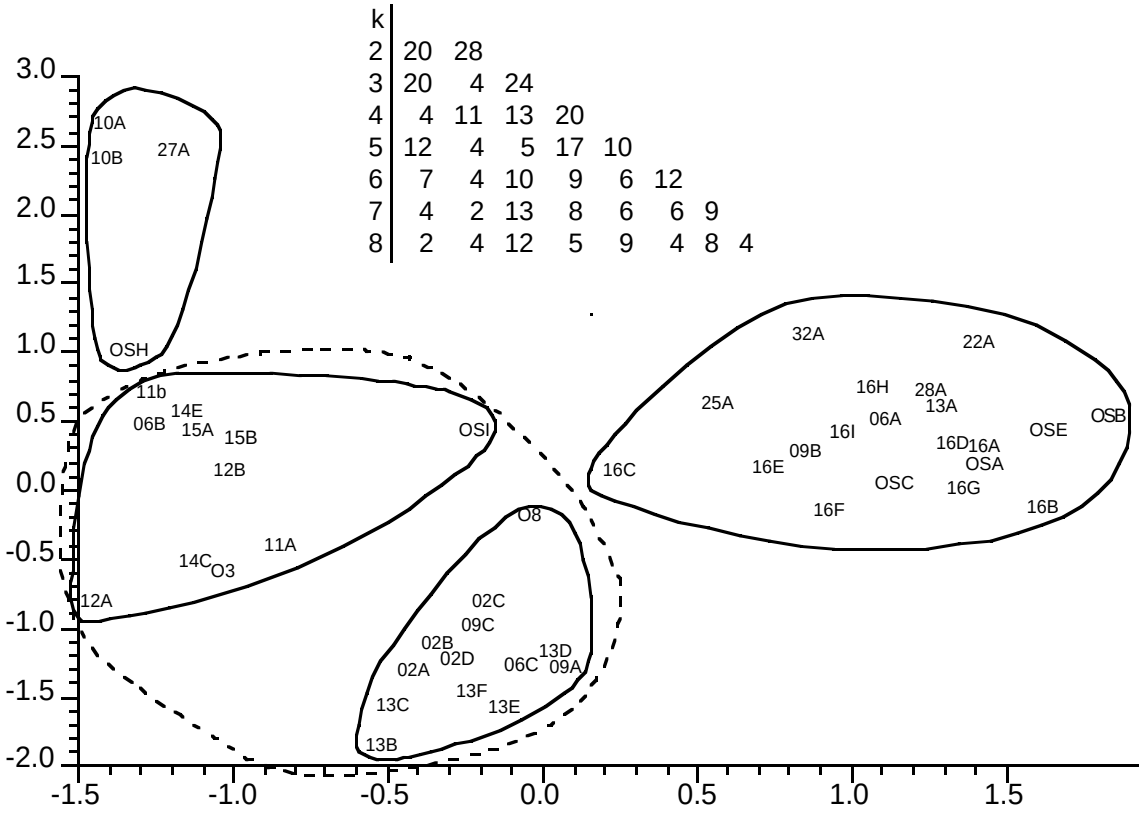


The analysis shows that the countries cluster together into two main groups along political lines: Group 1 contains the countries of the communist East Bloc (these data were collected during the Cold War). Group 2 contains countries of capitalist Western Europe. Moreover, there are two small clusters: (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.

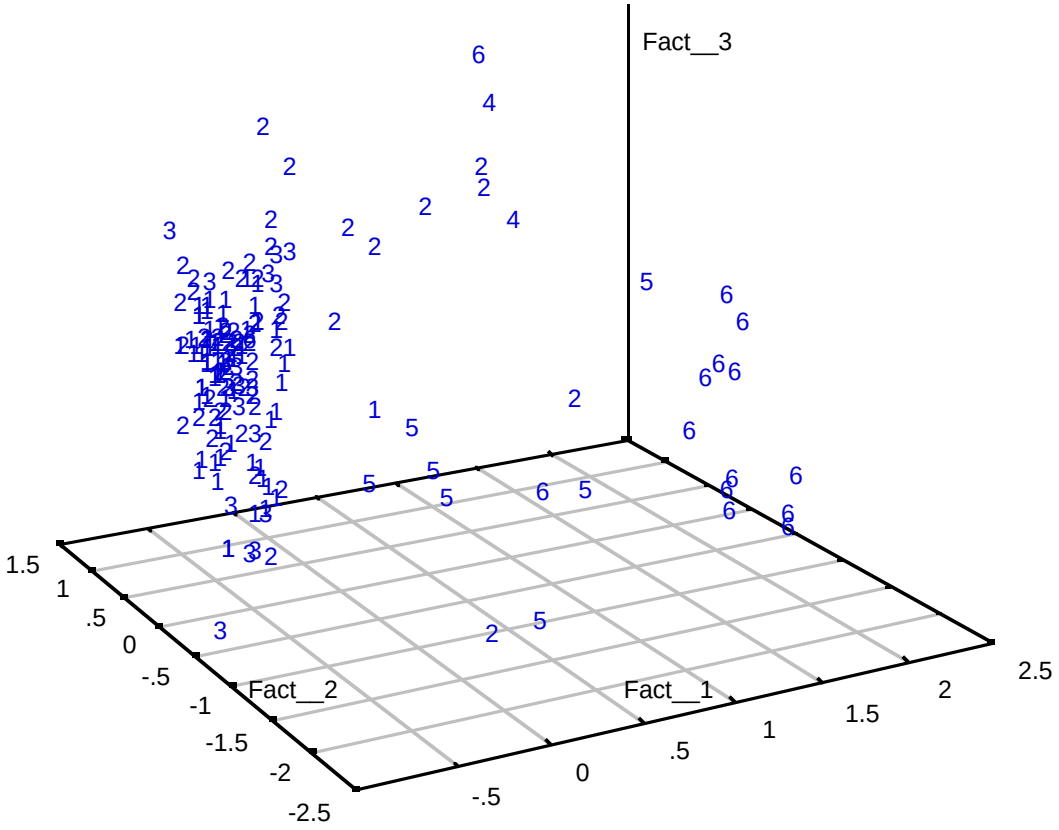
k	Criterion	Calinski	CH	Marriott	Si.Co.	D1	D2	$\tau 2$	C3
2	8.0443	13.73	13.73	0.3218	0.837	30.731	0.559	0.415	9.019
3	0.7983	18.57	18.57	0.0718	0.845	12.183	0.505	0.918	8.878
4	0.1518	27.60	27.60	0.0243	0.823	11.047	0.557	0.883	8.901
5	0.0614	22.21	22.21	0.0153	0.801	5.470	0.653	0.771	9.168
6	0.0285	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.

DetClus 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)

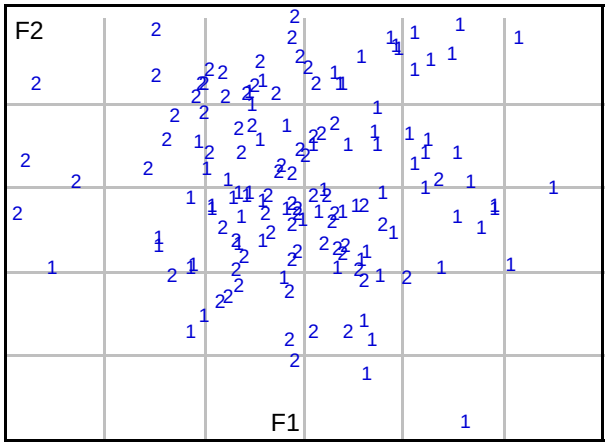
k									
2	50	164							
3	157	26	31						
4	161	31	2	20					
5	127	38	16	2	31				
6	23	18	18	2	31	122			
7	41	9	2	110	11	18	23		
8	3	11	23	2	33	18	115	9	
9	16	3	9	114	7	29	23	11	2
10	23	34	13	11	24	3	2	77	9	18

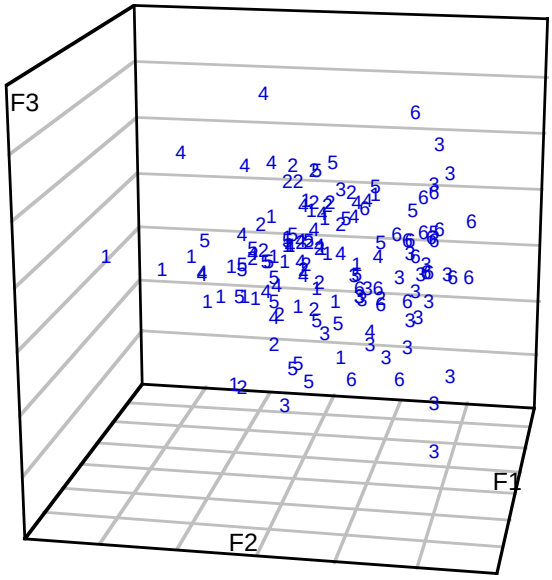


k	Criterion	Calinski	CH	Marriott	Si.Co.	D1	D2	τ2	C3
2	11.509	60.2	60.2	0.460	0.836	2.86	0.68	0.065	11.26
3	3.537	48.4	48.4	0.318	0.816	1.64	0.74	0.060	11.29
4	1.505	72.9	72.9	0.241	0.829	0.46	0.53	0.212	11.05
5	0.653	83.3	83.3	0.163	0.667	0.97	0.43	0.243	11.03
6	0.324	65.6	65.6	0.117	0.644	0.79	0.61	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- paroccipital 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**.



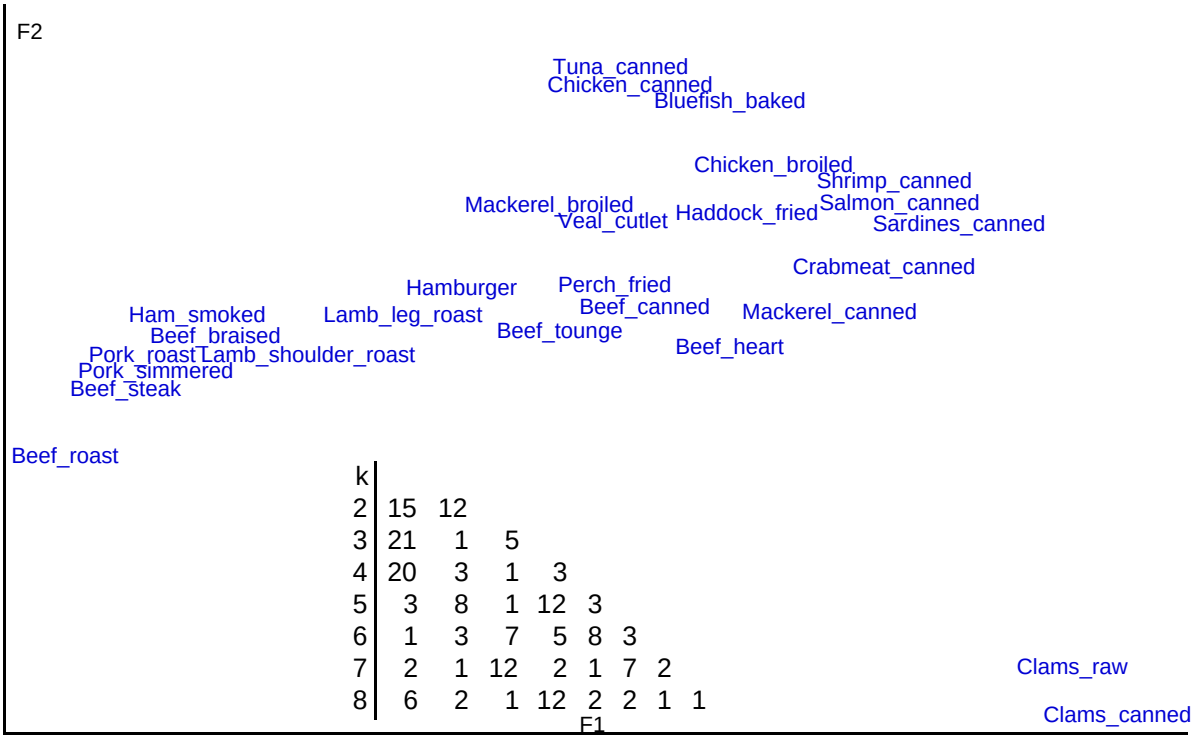


k	Criterion	Calinski	CH	C-index	Marriott	Si.Co.	D1	D2	D3	τ1	τ2	C3
2	21.442	13.96	13.96	0.1844	0.858	0.448	1.08	1.51	0.513	0.76	0.057	11.972
3	7.781	11.31	11.31	0.1067	0.700	0.375	1.24	1.17	0.704	0.73	0.047	11.903
4	3.748	10.79	10.79	0.0843	0.600	0.343	1.07	1.37	0.745	0.56	0.051	11.877
5	1.956	10.84	10.84	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).

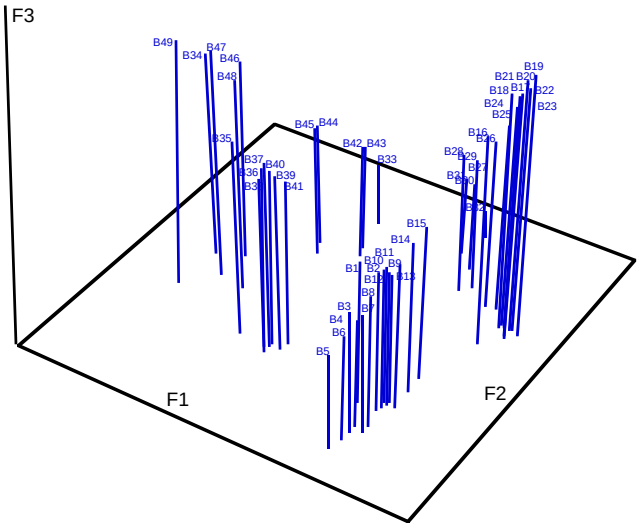


k							
2	15	12					
3	21	1	5				
4	20	3	1	3			
5	3	8	1	12	3		
6	1	3	7	5	8	3	
7	2	1	12	2	1	7	2
8	6	2	1	12	2	2	1	1

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 τ_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.



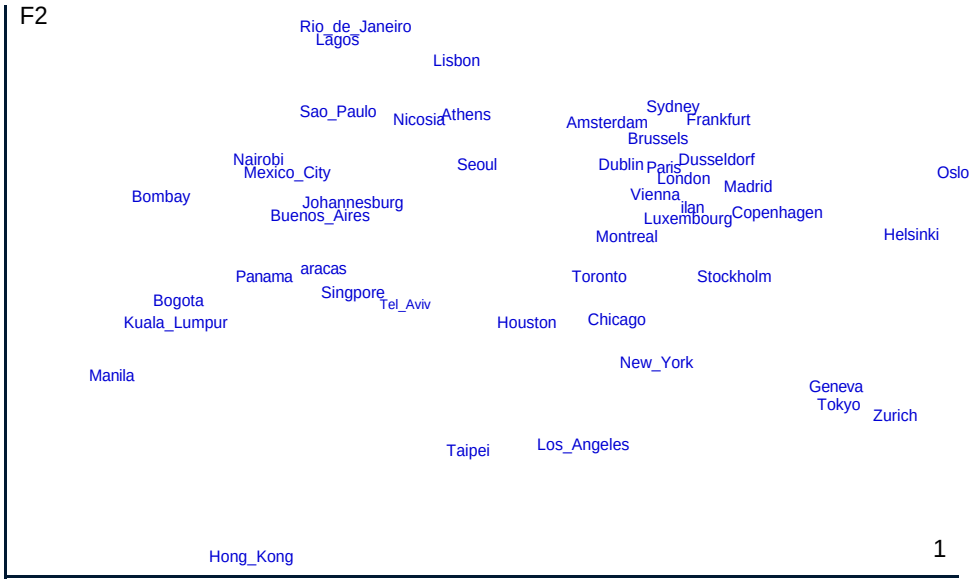
How to decide on the optimal number of clusters? Calinski-CH confirm *k=5* which is the best guess also for D_1, D_2, τ_2, and silhouette coefficient (this has a valley in the graph for *k=5*). The C^3 criterion has a major increment after *k=5* which can be interpreted as a sign of a good value for *k*.

k	Criterion	Calinski	CH	Marriott	Si.Co.	D1	D2	$\tau 2$	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 τ_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