Cluster Analysis: Overview

Article · December 2010

DOI: 10.1016/B978-0-08-044894-7.01315-4

2 authors:

Sabine Landau
King's College London
247 PUBLICATIONS **15,959** CITATIONS
SEE PROFILE

Irina Chis Ster
St George's, University of London
72 PUBLICATIONS **1,853** CITATIONS
SEE PROFILE

Cluster Analysis: Overview

S Landau, King's College London, London, UK
I Chis Ster, University College London, London, UK

The old idea of sorting similar things into groups underlies all aspects of human activity. For example, languages can be thought of as classification systems where words are used to describe types of events, objects, and people encountered. In addition to being a basic human conceptual activity, classification is fundamental to many branches of science with prominent examples being animal or plant taxonomies informing evolutionary theories in biology, or groupings of elements according to their chemical properties influencing research on the structure of the atom. In the social sciences, including education, it would typically be people that are to be grouped to identify patterns of behavior, achievement, etc.

Numerical methods aimed at discovering groups in data are referred to as cluster analysis. The groups can be sets of objects (individuals, countries, animals, chemical elements, etc.) or sets of variables. It also needs to be emphasized that cluster analysis is aimed at uncovering as-yet-unknown groups of objects; with analogous concepts being unsupervised pattern recognition or numerical taxonomy. In contrast, discriminant analysis or supervised pattern recognition aims to establish rules that classify objects into classes that are known *a priori* based on a set of observable characteristics. Finally, cluster analysis is an exploratory technique. Its primary aim is not to infer anything about population parameters as most statistical methods do – but rather to suggest groupings that might form the basis of future hypotheses to be investigated.

Cluster analysis techniques themselves can be broadly grouped into three classes labeled hierarchical clustering, optimization clustering, and model-based clustering. They operate either directly on a matrix of scores on a number of variables for a set of objects to be classified, or on a matrix of distances or similarities between the objects.

Proximity Measures

Ideally, clusters should be internally cohesive structures that are isolated from each other. To judge this adequacy criteria are needed that encapsulate the concepts of cluster homogeneity (cohesion) and separation (isolation). Such measures can be derived from a matrix of object distances or (dis)similarities; more generally referred to as proximities.

Let $\mathbf{X}$ denote the usual $n \times p$ multivariate data matrix containing the data values describing each object to be clustered,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \dots & \dots & x_{np} \end{bmatrix}$$

That is, the entry x_{ij} gives the value for the jth variable of the ith object. This is a two-mode matrix indicating that the rows and columns correspond to different things. For a raw data matrix containing only categorical variables, similarity measures are typically used. The similarity coefficient s_{ik} between objects i and k is one if both objects have identical values for all variables and zero if the two objects differ maximally for all variables. (The corresponding dissimilarity is then simply $d_{ik} = 1 - s_{ik}$.) In the binary case, similarity measures arise from a cross-classification of the counts of matches and mismatches of the p variables. **Table 1** shows some possible similarity measures. A multitude of measures has been proposed largely due to the uncertainty as to how to deal with the count of zero–zero matches (for a comprehensive list see Gower and Legendre (1986)). For categorical variables with more than two levels, similarities are typically calculated by allocating a score $s_{ik,j}$ of zero or one to each variable j, depending on whether the two objects i and k are the same on that variable and then simply averaging to give $s_{ik} = (1/p) \sum_j s_{ik,j}$. For a raw data matrix containing only continuous variables, proximities are typically expressed as distances (dissimilarity measures that fulfill the metric inequality); see **Table 2** for some possible measures. The Euclidean distance and the city block distance are appealing choices because of their geometric interpretation as physical distances between p-dimensional points with the latter traveled in rectilinear configuration.

As an example, consider applying the Euclidean distance measure to the protein consumption data in **Table 3** The consumption for 25 European countries is measured in grams of protein food per day. Clearly, if we were to calculate distances on the raw data, these would be dominated by the variables relating to products generally eaten in larger quantities, such as cereals. We therefore standardize the variables to unit variance before applying the distance measure, a technique sometimes referred to as autoscaling. **Table 4** shows the resulting distance matrix (a one-mode matrix).

Finally, there are a number of approaches for constructing proximities for mixed-mode data – data in

which some variables are continuous and others are categorical. A simple approach is to construct an appropriate proximity matrix for each variable type submatrix and then to combine the measures.

Returning to our objective of wanting to measure the adequacy of a clustering on the basis of proximities, there are two basic approaches to defining intergroup proximities: (1) define the proximity by a suitable summary of the proximities between individuals from either group, or (2) represent each group by a typical observation and measure the proximity between these centers. Suitable summaries for (1) are the nearest neighbor distance, the furthest neighbor distance, or the average distance. Under approach (2), groups might be represented by, for example, the mean over the objects for each variable (continuous variables only), the so-called centroid, or the object that has the smallest average dissimilarity to all other group members (the medoid).

Hierarchical Clustering

Members of this class of clustering techniques produce a nested sequence of partitions by merging (or dividing) clusters. At each stage of the sequence, the new partition is optimally merged (or divided) from the previous partition according to some adequacy criterion. In the end, the sequence of partitions ranges from a single cluster containing all the individuals to n clusters containing a single individual. The whole series of partitions is most conveniently described by a tree display called the dendrogram (see later examples). Agglomerative hierarchical clustering proceeds by a series of successive fusions of the n objects into groups. In contrast, divisive hierarchical methods divide the n individuals successively into finer groups. Divisive methods are not commonly used due to computational problems; for more details see Everitt *et al.* (2001).

A variety of agglomerative techniques exist reflecting the different ways in which intergroup dissimilarities can be defined. For example, merging two clusters when their nearest neighbor distance is minimal leads to an agglomerative procedure called single linkage. Similarly, the use of the furthest neighbor distance or the average distance leads to complete and average linkage, respectively. To describe the agglomerative process, complete linkage will be applied to the first five countries in **Table 3** using the Euclidean distances shown as the shaded area in **Table 4**. Initially, there are five clusters all containing a single country (partition 1). The first step is to combine the closest pair of countries. From **Table 4** this is seen to be Austria and Czechoslovakia (partition 2). The distances between Albania, Bulgaria, and Belgium and the cluster (Austria and Czechoslovakia) need to be evaluated next. Based on the furthest neighbor distance this is found as: $d_{\text{new country, cluster}} = \max(d_{\text{new country, Austria}}, d_{\text{new country, Czechoslovakia}})$. After these new distances are calculated, the smallest value is again used to decide which clusters should be merged; here Belgium and (Austria, Czechoslovakia) are fused (partition 3). The next stage involves evaluating the furthest neighbor distances between clusters [(Austria, Czechoslovakia), Belgium], Albania and Bulgaria. This leads to Albania and Bulgaria being combined (partition 4) and then in the final step [(Austria, Czechoslovakia), Belgium] and (Albania,

Table 1 Similarity measures for binary data

	Counts of binary outcomes for two individuals Individual i			
Individual *j*	Outcome	1	0	Total
	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

Similarity measures

Measure	*Formula*
Matching coefficient	$s_{ij} = (a+d)/(a+b+c+d)$
Jaccard coefficient	$s_{ij} = a/(a+b+c)$
Rogers and Tanimoto	$s_{ij} = (a+d)/[a+2(b+c)+d]$
Sokal and Sneath	$s_{ij} = a/[a+2(b+c)]$
Gower and Legendre I	$s_{ij} = (a+d)/\left[a+\frac{1}{2}(b+c)+d\right]$
Gower and Legendre II	$s_{ij} = a/\left[a+\frac{1}{2}(b+c)\right]$

Table 2 Distance measures for continuous data

Measure	*Formula*
Euclidean distance	$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik}-x_{jk})^2\right)^{1/2}$
City block distance	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik}-x_{jk}\rvert$
Minkowski distance	$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik}-x_{jk})^r\right)^{1/r} \quad (r \geq 1)$
Canberra distance	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik}=x_{jk}=0 \\ \sum_{k=1}^{p} \lvert x_{ik}-x_{jk}\rvert/(\lvert x_{ik}\rvert+\lvert x_{jk}\rvert) & \text{for } x_{ik}\neq 0 \text{ or } x_{jk}\neq 0 \end{cases}$

Table 3 Protein consumption in 25 European countries for nine food groups (grams per day)

Country	*Red meat*	*White meat*	*Eggs*	*Milk*	*Fish*	*Cereals*	*Starch*	*Nuts*	*Fruits and vegetables*
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

From Hand, D. J., Daly, F., McConway, K., Lunn, D., and Ostrowsky, E. (1994). *A Handbook of Small Data Sets*. London: Chapman and Hall.

Bulgaria) are fused to become a single cluster (partition 5). The fusion process is described graphically by the dendrogram in **Figure 1** with the y-axis showing the distance at which clusters are fused. (Note that the display is not unique; e.g., (Albania, Bulgaria) or [Belgium, (Austria, Czechoslovakia)] can be flipped.) A popular agglomerative method for continuous data which measures cluster adequacy by evaluating distances between cluster centroids is Wards method. The approach merges clusters if the sum of the squared dissimilarities to the cluster centroid (a measure of cohesion) is minimal across all possible merging choices. **Figure 2** shows results of applying single linkage (a), complete linkage (b) and Ward's method (c) to the full distance matrix from **Table 4**. Different distances produce different cluster solutions. A similarity may be observed between complete linkage and Ward's method, but the single linkage solution looks rather different.

Users need to be aware that partitions obtained by hierarchical clustering are irrevocable in the sense that once clusters have been combined in an agglomerative procedure they cannot be split up again. This means that not all possible partitions are evaluated which, while computationally convenient, may mean that an optimal partition is missed. Second, the method may not take account of the cluster structure properly. For example, single linkage is known to be prone to chaining, the tendency for new points to join the previous cluster in a chain-like fashion. This is what appears to have happened in **Figure 2(a)**. In contrast, complete linkage is known to produce compact clusters which may not always reflect the true data structure either. Efforts have been made to define properties that would be useful for hierarchical cluster methods. For example, point proportionality refers to replication of points not altering the boundaries of partitions. The monotone property, which states that the numerical values of the proximities are unimportant in that only their ranking matters for the clustering, might also be desirable. Single linkage and complete linkage, which otherwise are problematic, possess these properties; see Everitt *et al.* (2001).

Optimization Clustering

In contrast to hierarchical clustering methods optimization clustering methods aim to evaluate the adequacy of all possible partitions of a set of objects into k clusters. For the moment, we consider the number of clusters k to be known. The basic idea behind these methods is that associated with each partition of the n objects into the required number of groups k is an index $c(n, k)$, the value of which is to be optimized. Differences between the clustering methods in this class arise because of the variety of

Table 4 Euclidean distances for protein data after standardisation of variables to unit variance

	Alb	*Aus*	*Belg*	*Bulg*	*Czech*	*Denm*	*Germ*	*Finlan*	*France*	*Greece*	*Hungar*	*Ireland*	*Italy*	*Nether*	*Norwa*	*Poland*	*Portug*	*Roman*	*Spain*	*Swed*	*Switzer*	*UK*	*USSR*	*W Ger*	*Yugosl*
Alb	0																								
Aus	6.14	0																							
Belg	5.94	2.46	0																						
Bulg	2.76	4.9	5.24	0																					
Czech	5.14	2.14	2.22	3.96	0																				
Denm	6.64	3.02	2.54	6.04	3.36	0																			
Germ	6.4	2.58	2.12	5.4	1.88	2.76	0																		
Finlan	5.88	4.08	3.5	5.82	3.98	2.64	4.08	0																	
France	6.3	3.58	2.2	5.56	3.36	3.66	3.8	4.6	0																
Greece	4.26	5.16	4.7	3.76	4.88	5.6	5.62	5.5	4.54	0															
Hungar	4.68	3.28	4	3.34	2.76	5.04	3.68	5.4	4.98	4.12	0														
Ireland	6.76	2.74	1.66	6.22	3.16	2.82	3.04	3.24	3.16	5.7	4.82	0													
Italy	4.02	3.72	3.72	2.86	3.34	4.78	4.32	4.92	3.8	2.16	3.16	4.84	0												
Nether	6	1.12	2.24	5.16	2.2	2.54	2.54	3.38	3.4	5.16	3.5	2.34	3.92	0											
Norwa	5.46	3.88	2.96	5.28	3.52	2	3.28	2.06	3.92	4.62	4.9	3.6	4	3.36	0										
Poland	5.88	2.8	2.94	4.44	2.1	3.84	2.7	4.12	3.6	4.42	3.04	3.74	3.12	2.78	3.7	0									
Portug	6.62	6.52	5.66	6	5.52	5.86	5.26	6.5	5.66	4.78	5.7	7.06	4.66	6.36	4.8	4.82	0								
Roman	2.68	4.66	4.76	1.88	3.56	5.54	4.78	5.06	5.52	3.62	2.48	5.6	3.1	4.64	4.68	3.96	5.62	0							
Spain	5.56	4.88	4	4.84	4.14	5.12	4.08	5.48	4.46	3.1	3.88	5.28	2.88	4.86	4.16	3.4	2.94	4.24	0						
Swed	5.66	2.94	2.58	5.4	3.26	1.38	3.06	2.06	3.82	4.98	4.66	2.86	4.14	2.4	1.5	3.84	5.86	4.86	4.8	0					
Switzer	5.12	2.2	2.34	4.48	2.62	3.18	3.58	3.54	2.42	4.1	3.86	2.82	2.94	1.9	3.34	3.08	6.12	4.36	4.58	2.68	0				
UK	5.94	3.74	1.94	5.8	3.84	3.48	3.92	3.88	2.58	4.62	5.12	2.26	4.18	3.52	3.54	4.5	6.54	5.42	4.72	3.14	2.84	0			
USSR	4.34	4.16	3.16	3.84	2.72	4.16	3.42	3.46	4.24	4.12	3.42	3.9	3.56	3.88	3.26	2.92	5.08	2.76	3.62	3.78	3.8	4	0		
W Ger	6.36	1.64	1.42	5.62	2.18	2.4	1.9	3.66	2.94	5.36	3.9	1.8	4.14	1.28	3.3	3	6.14	5.1	4.6	2.46	2.28	2.9	3.9	0	
Yugosl	2.94	5.44	5.6	2	4.34	6.36	5.52	5.8	6.3	3.94	3.04	6.46	3.58	5.5	5.4	4.5	5.82	0.98	4.56	5.7	5.2	6.26	3.36	5.96	0

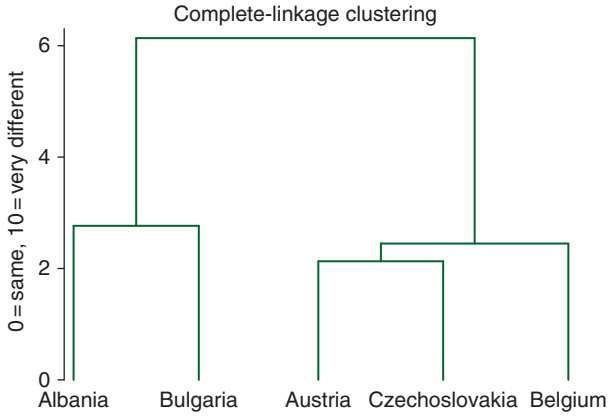


Figure 1 Dendrogram demonstrating complete linkage clustering of first five countries.

criteria that might be optimized. In addition, optimization clustering is a computing-intensive process since the number of necessary evaluations soon becomes very large as the number of clusters and/or the sample size increases. Thus, even with today's computers complete enumeration is not possible and consequently different algorithms have been suggested to optimize the same cluster criterion.

Cluster criteria can be defined on the basis of the proximity matrix or on the basis of the raw data matrix. When using the proximity matrix the criteria tend to define the lack of homogeneity for a single group (e.g., by the maximum dissimilarity between two objects from the group) and then the lack of homogeneity of a partition by suitably aggregating over the groups (e.g., by using a weighted sum over the groups). A number of algorithms for optimizing such cluster criteria were proposed by Kaufman and Rousseeuw (1990), with perhaps the most popular one being the partitioning around medoids (PAM) algorithm which minimizes criterion $c_{\mathrm{PAM}}(n,k) = \sum_{m=1}^{k}\sum_{i}^{n_m} d(\mathbf{x}_{im}, \mathrm{medoid}_m)$, the sum of the dissimilarities between the objects with p-dimensional data vectors $\mathbf{x}_{im}, m \in \{1,..,k\}, I \in \{1,...,n_m\}$ and their cluster medoids. As an example, consider 1993 data from 705 American colleges and universities on student intake, cost, and learning environment published by US News and World Report magazine in 1995. The data set is available at the math forum data collection website. Here, we have selected nine typical variables as shown in **Table 5** and use 705 colleges with complete information on these variables. The college sample is presented graphically in **Figure 3**. This raw data matrix was converted into a proximity matrix by calculating Euclidean distances. To overcome the unit of measurement problem, all variables were transformed so that they appeared to have symmetrical distributions (using ln and logit transformations) and then standardized to have sample mean 0 and standard deviation 1, before calculating distances. Application of the PAM algorithm provided the three clusters indicated in **Figure 3**. As PAM minimizes distances to the medoids, the latter represent the best cluster summary. **Table 6** shows that PAM cluster 1 consists of universities which enrol moderate numbers of students with a high proportion from the top 10% achievers, have the highest in-state tuition, smallest student-to-faculty ratio, and the highest proportion of graduates. Cluster 2 groups less selective universities that enrol the largest numbers of students, charge the least tuition fees, have a higher student-to-faculty ratio, and produce the smallest proportion of graduates. Cluster 3 contains less selective colleges with small enrolment numbers that charge moderate fees and have an average proportion of graduates.

Most cluster criteria derived from continuous data make use of a decomposition of the total dispersion matrix $\mathbf{T} = \sum_{m=1}^{k}\sum_{i=1}^{n_m}(\mathbf{x}_{im} - \bar{\mathbf{x}})(\mathbf{x}_{im} - \bar{\mathbf{x}})^{\mathrm{T}}$, where $\bar{\mathbf{x}}$ is the p-dimensional vector of overall sample means for each variable, into

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \qquad [1]$$

where $\mathbf{W} = \sum_{m=1}^{k}\sum_{i=1}^{n_m}(\mathbf{x}_{im} - \mathrm{centroid}_m)(\mathbf{x}_{im} - \mathrm{centroid}_m)^{\mathrm{T}}$ is the within-group dispersion matrix and $\mathbf{B} = \sum_{m=1}^{k} n_m (\mathrm{centroid}_m - \bar{\mathbf{x}})(\mathrm{centroid}_m - \bar{\mathbf{x}})^{\mathrm{T}}$ is the between-group dispersion matrix. Most popular in this class of cluster criteria is minimization of $c_{\mathrm{trace}}(n,k) = \mathrm{trace}\,(\mathbf{W})$. This is equivalent to minimizing the sum of the squared Euclidean distances between the objects and their centroids. Popularity is due to most software packages containing a k-means algorithm, which minimizes this criterion by iteratively reallocating an object to another group if the object is nearer (in terms of Euclidean distance to the centroid) to the new group than its own. K-means clustering of the colleges into three groups using the transformed data produced the solution in **Table 7**, presenting similar substantive results to PAM. Similar to PAM, k-means clustering minimized distances in some sense and is not scale invariant. A scale-invariant alternative cluster criterion is minimization of $c_{\mathrm{det}}(n,k) = \det(\mathbf{W})$. This cluster approach brings the practical benefit of not requiring the user to address the unit of measurement problem by standardization of variables or similar techniques. Both trace $(\mathbf{W})$ and det $(\mathbf{W})$ clustering have a tendency to generate clusters of roughly the same size (in terms of number of objects) and of similar shape and volume. Further criteria have been suggested to overcome this; see Everitt *et al.* (2001). Finally, a note of caution: as all the so-called hill-climbing algorithms aim to find a global optimum without having to evaluate all possible partitions, it is important to check convergence against a global optimum, for example, by rerunning the algorithm with a new set of starting values.

Model-Based Clustering

Most cluster analysis methods are essentially heuristic methods in the sense that they do not make explicit

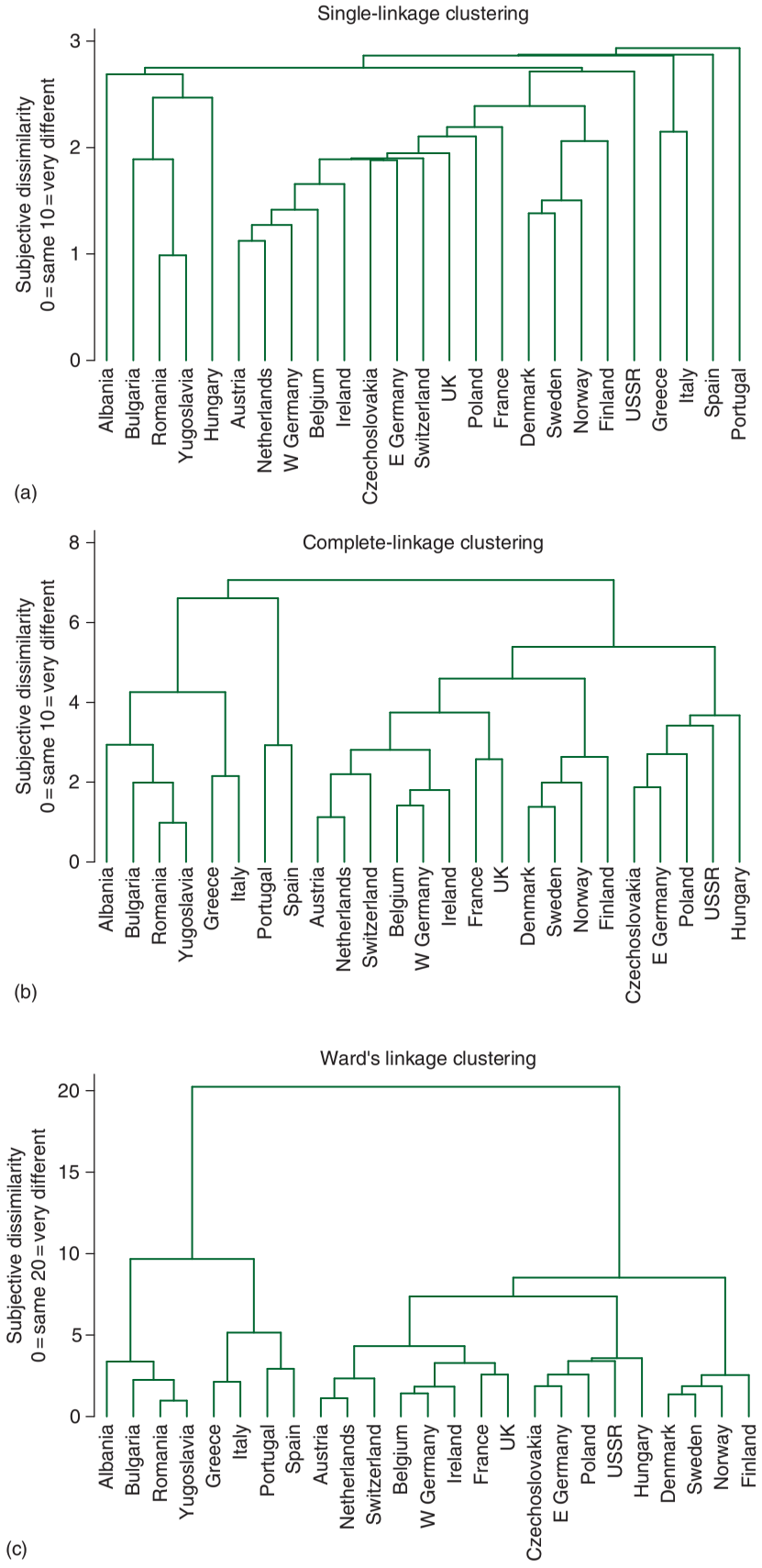


Figure 2 Dendograms showing partitions resulting from (a) single linkage, (b) complete linkage, and (c) Ward's method applied to protein consumption proximity matrix in **Table 4**.

Table 5 American colleges' data (first ten records only)

College name	*Rejection proportion*[a]	*Number enrolled*[b]	*Top 10%*[c]	*In-state tuition*[d]	*Additional fees*[e]	*Personal costs*[f]	*Number PhD faculty*[g]	*Ratio faculty students*[h]	*Percentage graduated*[i]
Alaska Pacific University	0.24	55	16	7560	130	1500	76	11.9	15
Chicago State University	0.52	777	12	1848	350	2400	47	15.6	18
Brewton-Parker College	0.14	1202	10	4371	130	2000	62	12.6	18
College of the Southwest	0.21	27	7	3120	125	500	24	14.3	20
Northeastern Illinois University	0.28	631	14	1902	236	2178	78	15.1	21
Mount Saint Clare College	0.13	95	16	9900	80	1200	32	13.6	21
Claflin College	0.42	499	21	4412	600	1000	69	16.9	21
Huron University	0.67	124	3	7260	330	1840	31	12.9	21
University of Colorado at Denver	0.53	261	30	1828	240	2138	89	18.1	24
Fayetteville State University	0.27	452	1	740	636	766	75	15.1	24

[a]Proportion of rejected applications.
[b]Number of students eventually enrolled.
[c]Percentage of new students from top 10% of high school class.
[d]State-specific tuition in $ per academic year.
[e]Additional costs: books and other materials for study in $ per academic year.
[f]Living and leisure costs in $ per academic year.
[g]Number of PhD staff.
[h]No students/no teachers ratio.
[i]Percentage of the students who graduated.

assumptions about the data-generating process. It is therefore impossible to infer from sample to population. Perhaps this presents no real difficulties to investigators involved in an initial exploration of their data where cluster analysis is only used to suggest hypothesis for future investigation. However, attempts have been made to develop a more acceptable statistical approach to the clustering problem, using what are known as finite mixture distributions (McLachlan and Peel, 2000).

Briefly, finite mixture densities are a family of probability density functions of the form

$$f(\mathbf{x}, \mathbf{p}, \boldsymbol{\theta}) = \sum_{m=1}^{k} p_m g_m(\mathbf{x}, \boldsymbol{\theta}_m) \qquad [2]$$

where $\mathbf{x}$ is a p-dimensional random variable and vectors $\mathbf{p}^\top = [p_1, p_2, \ldots, p_{k-1}]$ and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k]$ are parameter vectors. The $p_m \geq 0, m = 1, \ldots, k; \sum_m p_m = 1$ are known as mixing proportions and the $g_m(\mathbf{x}, \boldsymbol{\theta}_m), m = 1, \ldots, k$ are the component densities being parameterized by $\boldsymbol{\theta}_m$. The number of components forming the mixture is k.

Then, finite mixtures provide statistical models for cluster analysis if we assume that the objects within a cluster arise from one of k subpopulations with different multivariate distributions $g_m(\mathbf{x}, \boldsymbol{\theta}_m)$. The latter distributions may belong to the same family, but differ in the values they have for the parameters of the distributions or come from different families (e.g., Everitt and Bullmore, 1999). The mixing proportions and the parameters of the component densities can be estimated by maximum likelihood. This more formal statistical approach brings the advantage that one can develop cluster criteria whose optimization corresponds to maximizing the log-likelihood under a specified statistical model. This enables specification of the model assumptions under which a cluster criterion is expected to perform well.

Finally, having specified a suitable statistical model and estimated its parameters, the so-called model-based cluster analysis is typically performed by associating an object with a particular subpopulation (cluster) on the basis that this subpopulation maximizes the value of the estimated posterior probability:

$$\Pr(\text{object } i \text{ belongs to cluster } m \mid \mathbf{x}_i) = \frac{\hat{p}_m g_m\left(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_m\right)}{f\left(\mathbf{x}_i, \hat{\mathbf{p}}, \hat{\boldsymbol{\theta}}\right)} \quad [3]$$

over all possible subpopulations $m = 1, \ldots, k$.

A common finite mixture model is based on multivariate normal densities with different mean vectors and possibly different covariance matrices. It can be shown that for unconstrained component covariance matrices $\boldsymbol{\Sigma}_m$, $m = 1, \ldots, k$ maximization of the finite mixture likelihood is the same as minimizing criterion $c_{\text{unconstrained}}(n, k) = \prod_{m=1}^{k} \left[\det\left(\mathbf{W}_m / n_m^2\right)\right]^{n_m}$, where $\mathbf{W}_m$ denotes the (sample) dispersion matrix for the n_m-dimensional m-th subpopulation. If the covariances can be assumed to be the same across subpopulations (though remain unknown), that is, $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ for all m, the corresponding cluster criterion becomes $c_{\text{same cov}}(n, k) = \prod_{m=1}^{k} \left[\det\left(\mathbf{W} / n_m^2\right)\right]^{n_m}$. To explore the number of clusters present in the American colleges' data, **Figure 4** shows univariate distributions of the transformed data with overlaid densities. From this it appears that the number of visible subpopulations is 3 (see the three-peak distribution in **Figure 4(d)**). We therefore fitted a three-component multivariate normal mixture with constant (but unknown) component covariance matrices and used posterior probabilities to allocate cluster memberships. This leads to a cluster solution with similar interpretation to those obtained previously.

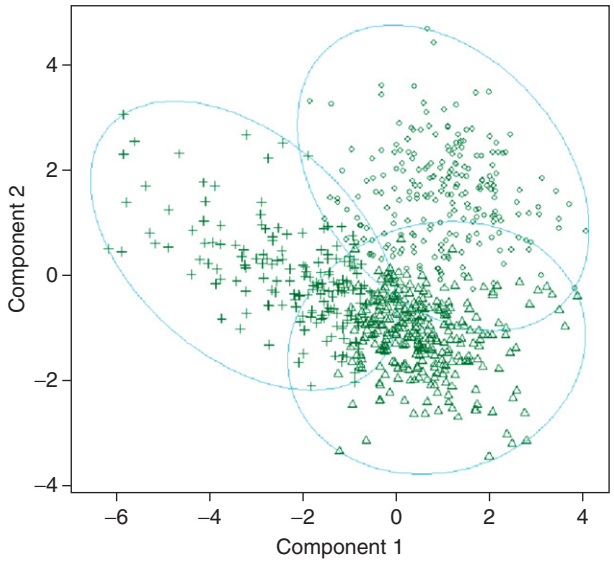


Figure 3 Scatter plot of colleges' data in the space of the first two principal components (symbols and ellipses indicate the three-group PAM solution).

Finite mixtures with multivariate normal distributions have been widely used because of their computational convenience. However, alternatives have been suggested; for example, multivariate t-distributions for groups of observations with longer tails than normal or atypical observations (McLachlan and Peel, 2000). For binary data multivariate Bernoulli densities which arise by assuming, that within each cluster, the binary variables are independent from each other (the so-called conditional independence assumption), are typically used. The latter multivariate component densities define the classical latent class model (Lazarsfeld and Henry, 1968) and have also been referred to as discrete mixture distributions. An interesting example is reported by Aitkin *et al.* (1981) who fitted latent class models to observations on 38 binary variables describing teaching behavior observations made on 468 teachers. Teachers were allocated to two classes by maximizing their posterior probabilities. **Table 8** summarizes the results by displaying response probabilities for each of the 38 items and each of the two classes. An obvious interpretation is a split into formal and informal teaching styles.

Table 6 Description of PAM three-cluster solution by medoids

	Student intake			*Affordability*			*Learning environment*		
Cluster (size)	*Rejection proportion*	*Number enrolled*	*Top 10%*	*In-state tuition*	*Additional fees*	*Personal costs*	*Proportion PhD (%)*	*Faculty/ student ratio*	*Proportion graduated (%)*
1 (178)	0.58	–0.38	1.02	1.11	–0.40	–0.83	1.16	0.99	0.92
2 (225)	0.08	1.22	–0.26	–0.17	0.67	0.37	0.27	–0.37	–0.40
3 (302)	−0.44	−0.75	−0.46	0.65	−0.37	−0.41	−0.57	0.19	0.16

Table 7 Description of *k*-means three-cluster solution by centroids (back-transformed to original scale)

	Student intake			*Affordability*			*Learning environment*		
Cluster (size)	*Rejection proportion*	*Number enrolled*	*Top 10%*	*In state tuition*	*Additional fees*	*Personal costs*	*Proportion PhD (%)*	*Faculty/ student ratio*	*Proportion graduated (%)*
1 (162)	0.36	786	48	15152	388	1076	87	0.1	82
2 (195)	0.28	1609	20	2497	529	1689	76	0.06	52
3 (347)	0.20	292	27	9910	252	1280	63	0.08	61

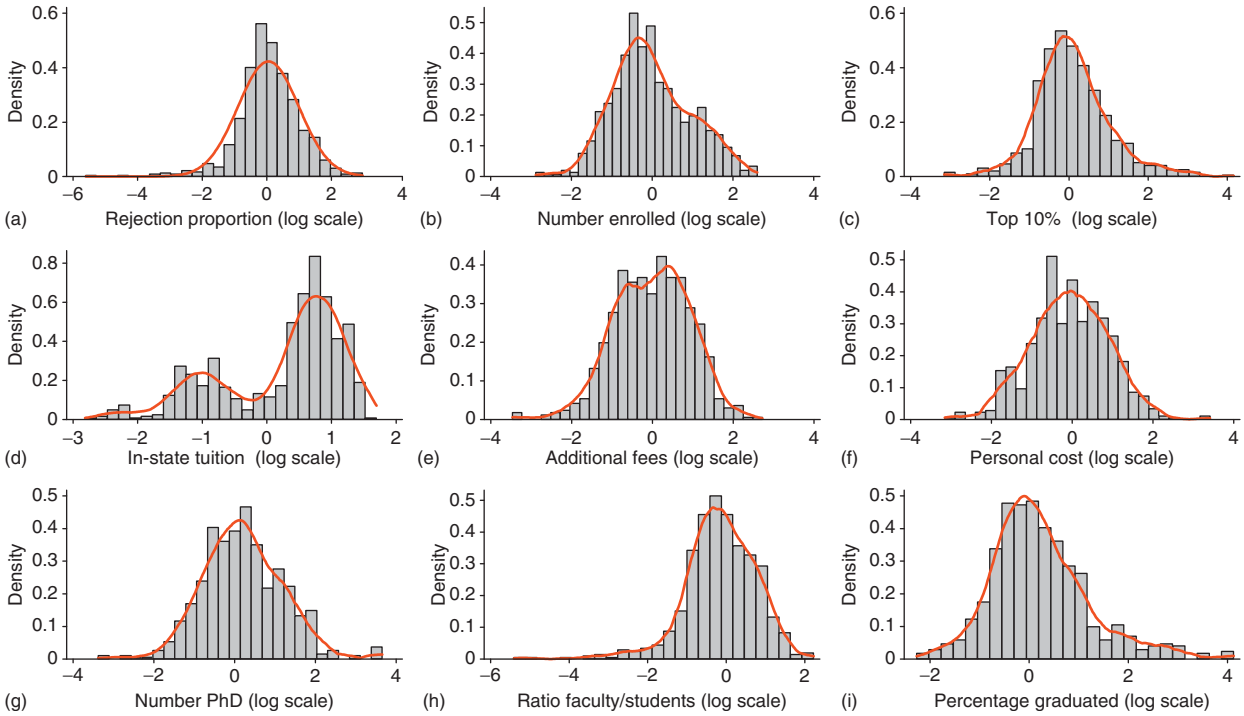


Figure 4 Histograms with overlaid densities for American college data.

Number of Clusters

We have so far studiously avoided the issue of choosing the number of clusters by assuming that an investigator is interested in the whole sequence of nested partitions or that the number of clusters is known *a priori*. However, in practice, one is often forced to decide the number of clusters from the information at hand.

A number of *ad hoc* procedures have been proposed. In the context of hierarchical clustering, selecting a partition is equivalent to cutting the corresponding dendrogram at a given height. This defines a partition such that clusters below that height are distant from each other by at least that amount. The appearance of the dendrogram can therefore informally suggest the number of clusters with large distances between fusion levels, suggesting the "best cut." For example, applying this rule to the dendrogram in **Figure 2(c)** might suggest five clusters. In the context of optimization, clustering scree plots, which plot the values of the optimized cluster criterion against the number of groups, are most popular. By definition, the criterion improves when the number of groups is increased – so the optimal number of groups is chosen as the level at which large changes in the criterion occur (the elbow in the plot). **Figure 5** shows the scree plot for *k*-means clustering of the college data; confirming that a partition into three clusters was appropriate.

More formal techniques exist which try to overcome the problem of subjectivity. Thirty such methods were reviewed by Milligan and Cooper (1985) in the 1980s and, more recently, 15 indices for high-dimensional binary data were assessed by Dimitriadou *et al.* (2002). Both studies assess the ability of formal/automated methods to detect the correct number of clusters in series of simulated data sets. Based on these simulations techniques introduced by Calinski and Harabasz (1974) and Duda and Hart (1973) are recommended for continuous data, while the index suggested by Ratkowsky and Lance (1978) was the overall best performer for binary data.

More recently, progress on defining formal rules for comparing the quality of different cluster solutions has been made in the context of model-based clustering. As, in this context, cluster memberships are determined by optimizing likelihoods of competing nested models; the latter can be compared to inform the choice of the number of groups. There are problems with the conventional likelihood ratio test as some parameters from the null distribution are on the edge of the parameter space. However, suggestions have been made to overcome this, with perhaps the most practical one being the use of information criteria such as the Bayesian information criterion (BIC) or Akaike's information criterion (AIC) (Burnham and Anderson, 2002). As an example, we calculated BIC for a number of models for the college data which differed in the numbers of clusters (from 1 to 5) and in the parametrization of the

Table 8 Estimated probabilities (in %) of responding yes to items for two classes of teachers

		Class 1	*Class 2*
1	Students have choice in where to sit	22	43
2	Students sit in groups of three or more	60	87
3	Students allocated to sitting by ability	35	23
4	Students stay in same seats for most of day	91	63
5	Students not allowed freedom of movement in classroom	97	54
6	Students not allowed to talk freely	89	48
7	Students expect to ask permission to leave room	97	76
8	Students expected to be quiet	82	42
9	Monitors appointed for special jobs	85	67
10	Students taken out of school regularly	32	60
11	Timetable used for organizing work	90	66
12	Use own materials rather then text books	19	49
13	Students expected to know tables by heart	92	76
14	Students asked to find own reference material	29	37
15	Students given homework regularly	35	22
16	Teacher talks to whole class	71	44
17	Students work in groups on teacher class	29	42
18	Students work in groups on work of their own choice	14	46
19	Students work individually on teacher tasks	55	37
20	Students work individually on work of their own choice	28	50
21	Explore concepts in number work	18	55
22	Encourage fluency in English language even if inaccurate	87	94
23	Students work marked or graded	43	14
24	Spelling and grammatical errors corrected	84	68
25	Stars given to students who produce best work	57	29
26	Arithmetic test given at least once a week	59	38
27	Spelling test given at least once a week	73	51
28	End-of-term tests given	66	44
29	Many students who create discipline problems	09	09
30	Verbal reproof sufficient	97	95
31	Discipline: extra work given	70	53
32	Smack	65	42
33	Withdrawal of privileges	86	77
34	Send to headteacher	24	17
35	Send out of room	19	15
36	Emphasis on separate subject teaching	85	50
37	Emphasis on esthetic subject teaching	55	63
38	Emphasis on integrated subject teaching	22	65
Percentage of teachers attributed to class		54	46

covariance matrix. The latter were chosen to reflect different constraints displaying natural geometric features which can be derived from its spectral decomposition. Clusters exhibiting similar orientation, shape, or volume, satisfying two or all these restrictions, may be desirable within a particular classification context (Bensmail *et al*, 1997). A range of these restrictions and their results for model-based selection are displayed in **Figure 6**: for instance, EVI

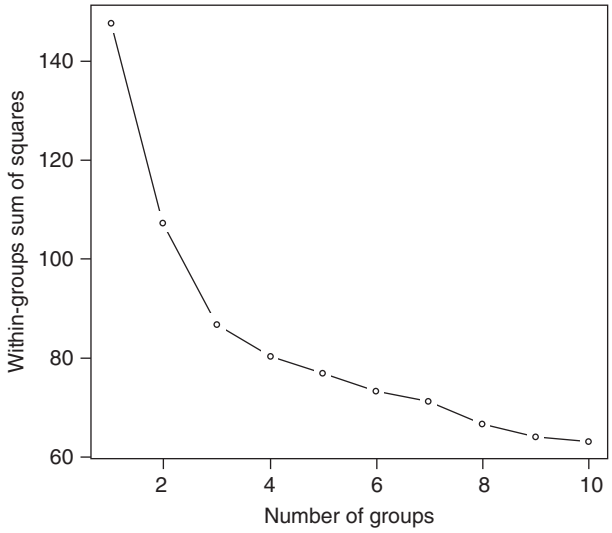
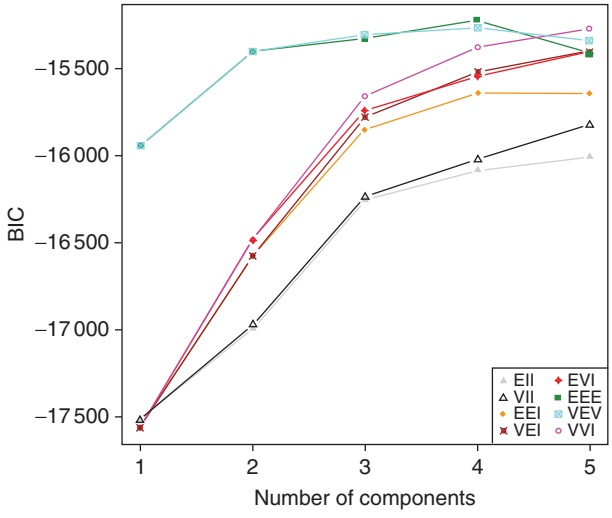


Figure 5 Scree plot for *k*-means clustering of colleges' data.



Distributional shape implied by covariance matrix parameterization (see Fraley and Raftery, 2003)

EII: spherical, equal volume
EVI: diagonal, equal volume, varying shape
VII: spherical, unequal volume
EEE: ellipsoidal, equal volume, shape, and orientation
EEI: diagonal, equal volume and shape
VEV: ellipsoidal, equal shape
VEI: diagonal, varying volume, equal shape
VVI: diagonal, varying volume and shape

Figure 6 Model selection by maximizing BIC. The best model is EEE with four clusters.

indicates that the clusters are constrained to having the same diagonal orientation and equal volume with their shape allowed to vary. Other symbols in **Figure 6** have analogous interpretations. The results show that the EEE criterion (which assumes the same covariance matrix and is equivalent to requiring similar orientation, volume, and shape for all clusters) performs well with three or four clusters.

Finally, it needs to be emphasized that a clustering algorithm will allocate objects into a prescribed number of groups, irrespective of whether there is any true clustering in the data. Many *ad hoc* methods for comparing the number of groups do not allow the consideration of the simplest clustering solution – the one-group solution. However, when empirically determining the number of clusters it is essential that the one-group solution be considered. Rules derived from maximum likelihood theory will allow this. An alternative approach is the GAP statistic (Tibshirani *et al.*, 2001), which compares the quality of cluster solutions for different numbers of groups based on a given (heuristic) cluster criterion.

Conclusion

The general steps involved in a cluster analysis are:

1. definition of the data matrix (including choice, weighting, or standardization of variables);
2. calculation of the proximity matrix;
3. choice of cluster method (to generate a single or a sequence of partitions);
4. decision regarding the number of clusters (for partitions); and
5. validity checks.

We have talked about steps 1–4 in previous sections but not much has yet been said about the final stage – checking the validity of a cluster solution. There are basically two approaches: internal checks and external checks. Internal checks are aimed at establishing cluster isolation and cohesion or demonstrating robustness of the solution under small changes of method (change of proximity measure, optimization criterion, starting values, etc.) or data set used (splitting data into subsamples, adding an error term, etc.). External checks attempt to establish agreement with a gold standard (if one exists) or with some other yet-unused variable/classification that one would theoretically expect to be associated with the solution. For more information on methods for validation checking see Everitt *et al.* (2001).

There are a number of related techniques not yet mentioned. Constrained clustering imposes restrictions on the possible cluster solutions in order to maintain external features; for example, spatial contiguity. Many applications, particularly in psychology and the social sciences, require overlapping clusters; that is, an object is allowed to be a member of more than one cluster at the same time – for example, a member of several overlapping social networks. Under some circumstances, direct data clustering, which aims to cluster the (two-mode) data matrix into sets of similar objects and variables, can be applied. Finally, neural network techniques, such as Kohonen's self-organizing map, aimed at unsupervised learning can be considered a cluster method.

Nowadays, carrying out cluster analyses is relatively straightforward. Most general-purpose statistical packages contain procedures for hierarchical and optimization clustering. (The analyses presented here were generated in Stata and R.) Routines for model-based clustering are available in some general-purpose packages (e.g., mclust in R, Fraley and Raftery, 2003), specialized latent classes and finite mixtures programs (e.g., LatentGOLD and MIXMOD), or modeling packages such as Mplus. In addition, there are a number of packages solely devoted to cluster analyses (e.g., Clustan).

In summary, clustering a set of objects can potentially be very useful. However, care needs to be taken to avoid producing misleading results. Researchers do well to remember that cluster analysis is an exploratory technique rather than an inferential method.

See also: Discrimination and Classification; Latent Class Models.

Bibliography

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society A* **144**, 419–448.

Bensmail, H., Celeux, G., Raftery, E. A., and Robert, C. P. (1997). Inference in model based cluster analysis. *Statistics and Computing* **7**, 1–10.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information–Theoretical Approach,* 2nd edn. New York: Springer.

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1–27.

Dimitriadou, E., Dolničar, S., and Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika* **67**, 137–160.

Duda, R. O. and Hart, P. E. (1973). *Pattern Analysis and Scene Analysis*. New York: Wiley.

Everitt, B. S. and Bullmore, E. T. (1999). Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping* **7**, 1–14.

Hand, D. J., Daly, F., McConway, K., Lunn, D., and Ostrowsky, E. (1994). *A Handbook of Small Data Sets*. London: Chapman and Hall.

Fraley, C. and Raftery, A. E. (2003). Enhanced software for model-based clustering, discriminant analysis, and density estimation: MCLUST. *Journal of Classification* **20**, 263–286.

Gower, J. C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* **5**, 5–48.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data, an Introduction to Cluster Analysis*. New York: Wiley.

Lazarsfeld, P. L. and Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.
Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159–179.
Ratkowsky, D. A. and Lance, G. N. (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal* **10**, 115–117.
Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, B* **63**, 411–423.

Further Reading

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster Analysis,* 4th edn. London: Arnold.

Relevant Website

http://mathforum.org – the Math Forum Data Collection.