CHAPTER

Cluster Validity

16

16.1 INTRODUCTION

A common characteristic of the majority of the clustering algorithms, discussed in the previous chapters, is that they *impose* a clustering structure on the data set X, even though X may not possess such a structure. In the latter case, the results produced after the application of a clustering algorithm on X are not indicative of the structure of X. In other words, *cluster analysis is not a panacea*. That is, we must have an indication that the vectors of X form clusters before we apply a clustering algorithm. The problem of verifying whether X possesses a clustering structure, without identifying it explicitly, is known as *clustering tendency* and is discussed at the end of the chapter.

Let us now assume that X possesses a clustering structure and we want to unravel it. A different kind of problem is encountered now. Recall that all the clustering algorithms require knowledge of the values of specific parameters and, in addition, some of them impose restrictions on the shape of the clusters (e.g., compact, hyperellipsoidal). As already shown in the previous chapters, poor estimation of these parameters and inappropriate restrictions on the shape of the clusters (wherever such restrictions are required) may lead to incorrect conclusions about the clustering structure of X. Thus, the need for further evaluation of the results of a clustering algorithm is apparent.

In this chapter, we discuss methods suitable for quantitative evaluation of the results of a clustering algorithm. This task is known under the general term *cluster validity*. However, it must be emphasized that the results obtained by these methods are *only* tools at the disposal of the expert in order to evaluate the resulting clustering.

Let $\mathcal{C}$ denote the clustering structure resulting from the application of a clustering algorithm on X. This may be a hierarchy of clusterings, as is the case with the hierarchical algorithms, or a single clustering, as happens with all the other algorithms discussed in the previous chapters. Cluster validity can be approached in three possible directions. First, we may evaluate $\mathcal{C}$ in terms of an independently drawn structure, which is imposed on X *a priori* and reflects our intuition about the clustering structure of X. The criteria used for the evaluation of this kind are called *external criteria*. In addition, external criteria may be used to measure the degree to which the available data confirm a prespecified structure, without applying any clustering algorithm to X. Second, we may evaluate $\mathcal{C}$ in terms of quantities that involve the vectors of X themselves, for example, the proximity matrix. The criteria used for this kind of evaluation are called *internal criteria*. Finally, we may evaluate $\mathcal{C}$ by comparing it with other clustering structures, resulting from the application of the same clustering algorithm, but with different parameter values, or of other clustering algorithms to X. Criteria of this kind are called *relative criteria*.

The cluster validation methods based on external or internal criteria rely on statistical hypothesis testing, which was introduced in Chapter 5. The following section contains some additional definitions to be used in this chapter.

16.2 HYPOTHESIS TESTING REVISITED

Let H_0 and H_1 be the null and alternative hypotheses, respectively,

$$H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

Also let $\bar{D}_\rho$ be the critical interval corresponding to significance level ρ of a test statistic q, and Θ_1 the set of all possible values that $\boldsymbol{\theta}$ may take under hypothesis H_1. The *power function* of the test is defined as

$$W(\boldsymbol{\theta}) = P(q \in \bar{D}_\rho | \boldsymbol{\theta} \in \Theta_1) \tag{16.1}$$

For a specific $\boldsymbol{\theta} \in \Theta_1$, $W(\boldsymbol{\theta})$ is known as the *test power under the alternative* $\boldsymbol{\theta}$. In words, $W(\boldsymbol{\theta})$ is the probability that q lies in the critical region when the value of the parameter vector is $\boldsymbol{\theta}$. This is the probability of making the correct decision when H_0 is rejected. The power function can be used for the comparison of two different statistical tests. The test whose power under the alternative hypotheses is greater is always preferred.

There are two types of errors associated with a statistical test.

- Suppose that H_0 is true. If $q(\boldsymbol{x}) \in \bar{D}_\rho$, H_0 will be rejected even if it is true. This is called a type I error. The probability of such an error is ρ. The probability of accepting H_0 when it is true is $1 - \rho$.
- Suppose that H_0 is false. If $q(\boldsymbol{x}) \notin \bar{D}_\rho$, H_0 will be accepted even if it is false. This is called a type II error. The probability of such an error is $1 - W(\boldsymbol{\theta})$, and it depends on the specific value of $\boldsymbol{\theta}$.

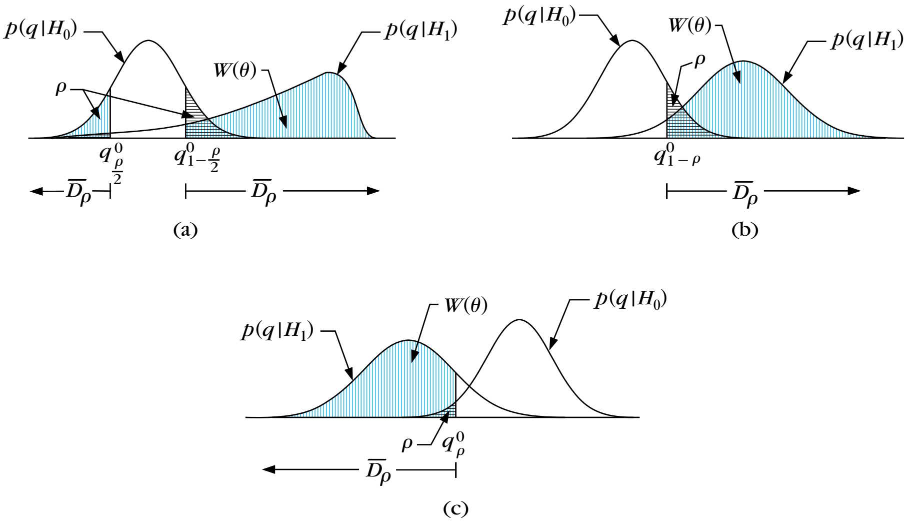


FIGURE 16.1

Critical regions of (a) A two-tailed test, (b) A right-tailed test, and (c) A left-tailed test. q^0_a is the a percentile of q under H_0.[1]

In practice, the final decision to reject or accept H_0 is based partially on the preceding statements as well as on other factors, such as the cost of a wrong decision. Thus, the terms "accept" and "reject" H_0 must be interpreted accordingly.

The probability density function (pdf) of the statistic q, under H_0, for most of the statistics used in practice has a single maximum and the $\bar{D}_\rho$ region is either a half-line or the union of two half-lines. These assumptions have also been adopted here. Figure 16.1 shows the three possible cases for $\bar{D}_\rho$. In the first case, $\bar{D}_\rho$ is the union of two half-lines. Such a test is known as a ***two-tailed statistical test***. The other two tests are called *one-tailed* statistical tests, because $\bar{D}_\rho$ consists of a single half-line. Figure 16.1a is an example of a two-tailed statistical test[2] and Figures 16.1b and 16.1c are examples of a right- and a left-tailed test, respectively.

In many practical cases the exact form of the pdf of a statistic q, under a given hypothesis, is not available and it is difficult to obtain. In the sequel, we discuss two methods for estimating pdf's via simulations.

- *Monte Carlo techniques* [Shre 64, Sobo 84] rely on simulating the process at hand using a sufficient number of computer-generated data. For each of the, say r, data sets, X_i, we compute the value of q, denoted by q_i, and then we construct the corresponding histogram of these values. The unknown pdf can

[1] The a percentile of q is the smallest number q_a such that $a = P(q \leq q_a)$.
[2] More general versions of a two-tailed statistical test are also possible (e.g., [Papo 91]).

then be approximated by this histogram. Assume now that q corresponds to a right-tailed statistical test and a histogram is constructed using r values of q corresponding to the r data sets. For a given data set, if q is the corresponding value of the statistic, rejection (acceptance) of the null hypothesis is done on the basis of

$$\text{Reject (accept) } H_0 \text{ if } q \text{ is greater (smaller) than } (1-\rho)r \text{ of the } q_i \text{ values} \tag{16.2}$$

For a left-tailed test, rejection or acceptance of the null hypothesis is done on the basis of

$$\text{Reject (accept) } H_0 \text{ if } q \text{ is smaller (greater) than } \rho r \text{ of the } q_i \text{ values} \tag{16.3}$$

Finally, for a two-tailed test we have

$$\begin{aligned}\text{Accept } H_0 \text{ if q is greater than } (\rho/2)\, r \text{ of the } q_i \text{ values and}\\ \text{less than } (1-\rho/2)\, r \text{ of the } q_i \text{ values}\end{aligned} \tag{16.4}$$

- *Bootstrapping techniques* constitute an alternative way to cope with a limited amount of data. The idea here is to parameterize the unknown pdf in terms of an unknown parameter. To cope with the limited amount of data and in order to improve the accuracy of the estimate of the unknown pdf parameter, several "fake" data sets $X_1, \ldots, X_r$ are created by sampling X with replacement, as discussed in Chapter 10.

  Typically, good estimates are obtained if r is between 100 and 200. For a more detailed discussion and applications of the bootstrapping techniques see, for example, [Diac 83, Efro 79, Jain 87a, Jain 87b].

16.3 HYPOTHESIS TESTING IN CLUSTER VALIDITY

In this framework, the null hypothesis H_0 will be expressed in a slightly different way. This is because our major concern is not to test a parameter against a specific value. In contrast, our concern here is to test whether the data of X possess a "random" structure or not. *Thus, in this case, the null hypothesis H_0 should be a statement of randomness concerning the structure of X.* Thus, our goal is now twofold.

- First, we must generate a reference *data population under the random hypothesis*, that is, a data population that models a random structure.
- Second, we must define an appropriate statistic, whose values are indicative of the structure of a data set, and compare the value that results from our data set X against the value obtained from the reference (random) population.

There are three different ways to generate the reference population under the null (randomness) hypothesis, each being appropriate for different situations.

- *Random position hypothesis.* This hypothesis is appropriate for ratio data. It *requires* that *"All the arrangements of N vectors in a specific region of the l-dimensional space are equally likely to occur."* Such regions may be the H_l hypercube or the l-dimensional hypersphere. One way to produce such an arrangement is to insert each point randomly in this region of the l-dimensional space, according to the uniform distribution. The random position hypothesis can be used with either external or internal criteria.

  - *Internal criteria.* In this case, the statistic q is defined so as to measure the degree to which a clustering structure, produced by a clustering algorithm, matches the proximity matrix of the corresponding data set. Let X_i be a set of N vectors generated according to the random position hypothesis and P_i be the corresponding proximity matrix. In the sequel, we apply the same clustering algorithm to each X_i and to our data set X and let $\mathcal{C}_i$ and $\mathcal{C}$ be the resulting clustering structures, respectively. For each case, the value of the statistic q is computed. The random hypothesis, H_0, is then rejected if the value q, resulting from X lies in the critical interval $\bar{D}_\rho$ of the statistic pdf of the reference population (i.e., under H_0), that is, if q is unusually small or large.

  - *External criteria.* The statistic q is defined so as to measure the degree of correspondence between a *prespecified structure* $\mathcal{P}$ imposed on X and the clustering that results after the application of a specific clustering algorithm to X. Then, the value of q corresponding to the clustering $\mathcal{C}$ resulting from the data set X is tested against the q_i's, corresponding to the clusterings resulting from the reference population generated under the random position hypothesis. Once more, the random hypothesis is rejected if q is unusually large or small.

- *Random graph hypothesis.* It is usually adopted when only internal information (i.e., information that concerns only the vectors themselves or their relationships) is available. It is appropriate when ordinal proximities between vectors are used. Before we proceed, let us define the ordinal, or rank order, $N \times N$ matrix A as a symmetric matrix with zero diagonal elements (provided that dissimilarity measures are used) and with its upper diagonal elements being integers in the range $[1, N(N-1)/2]$. The entry $A(i,j)$ of A provides only qualitative information about the dissimilarity between the corresponding vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. If, for example, $A(2,3) = 3$ and $A(2,5) = 5$, we can only conclude that $\boldsymbol{x}_2$ is more similar to $\boldsymbol{x}_3$ than $\boldsymbol{x}_5$. That is, in this context, comparing dissimilarities is meaningless (recall the comments made in Chapter 11, concerning ordinal type data).

  Let A_i be an $N \times N$ rank order proximity matrix with no ties; that is, all entries in the upper diagonal are different from each other. Under the random graph hypothesis, the reference population consists of such matrices A_i each one generated by inserting randomly the integers in the range

$[1, N(N-1)/2]$, in its upper diagonal entries. Let P be the ordinal proximity matrix associated with the given data set X and $\mathcal{C}$ be the clustering structure produced by the application of a specific algorithm to P. Finally, let $\mathcal{C}_i$ be the clustering structure produced when the same algorithm is applied to A_i. We may now proceed as in the previous case and define a statistic q that measures the agreement between a rank order (proximity) matrix and the corresponding clustering structure. If the value of q, corresponding to P and $\mathcal{C}$, is unusually large or small, the random hypothesis is rejected.

It must be emphasized that the random graph hypothesis is not appropriate for ratio-scaled data. Let us take, for example, the case where the Euclidean distance is in use and $l \leq N-2$ and consider the points $x_1 = 0$, $x_2 = 1$, $x_3 = 3$ on the real line. It is clear that the distance between x_1 and x_3 cannot be smaller than the distance between x_2 and x_3. That is, the matrix $A = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$ is not a valid proximity matrix for these ratio-scaled data.

- *Random label hypothesis.* Let us consider all possible partitions, $\mathcal{P}'$, of X into m groups. Each partition may be defined in terms of a mapping g from X to $\{1, \ldots, m\}$. The random label hypothesis assumes that *all* possible mappings are *equally likely*. The statistic q can be defined so as to measure the degree to which information inherent in the data set X, such as the proximity matrix P, matches a specific partition. The statistic q is then used to test the degree of match between P and an externally imposed partition $\mathcal{P}$, against the q_i's corresponding to the random partitions generated under the random label hypothesis. Once more, H_0 is then rejected if q is unusually large or small.

In the sequel, we give a number of statistic indices appropriate for external and, then, for internal criteria.

16.3.1 External Criteria

External criteria are used either (a) for the comparison of a clustering structure $\mathcal{C}$, produced by a clustering algorithm, with a partition $\mathcal{P}$ of X drawn independently from $\mathcal{C}$ or (b) for measuring the degree of agreement between a predetermined partition $\mathcal{P}$ and the proximity matrix of X, P.

Comparison of $\mathcal{P}$ with a Clustering $\mathcal{C}$

In this case, $\mathcal{C}$ may be either a specific hierarchy of clusterings or a specific clustering. The latter may be produced either by cutting the dendrogram produced by a hierarchical algorithm at a given level (see Chapter 13) or by any other algorithm discussed in the previous chapters. However, a prespecified hierarchy of partitions is rarely available in practice. Thus, the problem of validating hierarchies of clusterings is of limited practical interest.

In the sequel, we consider the validation task concerning a clustering, $\mathcal{C}$, resulting from a specific clustering algorithm, in terms of an independently drawn partition $\mathcal{P}$ of X. Let $\mathcal{C} = \{C_1, \ldots, C_m\}$ and $\mathcal{P} = \{P_1, \ldots, P_s\}$. Note that the number of clusters in $\mathcal{C}$ need not be the same as the number of groups in $\mathcal{P}$. Our goal is to define appropriate statistical indices to be used for the hypothesis test.

Let n_{ij} denote the number of vectors that belong to C_i and P_j simultaneously. Also let $n_i^C = \sum_{j=1}^{s} n_{ij}$; that is, n_i^C is the number of vectors that belong to C_i. Similarly, we define the number of vectors that belong to P_j as $n_j^P = \sum_{i=1}^{m} n_{ij}$.

Consider a pair of vectors $(\boldsymbol{x}_v, \boldsymbol{x}_u)$. We refer to it as (a) *SS* if both vectors belong to the same cluster in $\mathcal{C}$ and to the same group in $\mathcal{P}$, (b) *DD* if both vectors belong to different clusters in $\mathcal{C}$ and to different groups in $\mathcal{P}$, (c) *SD* if the vectors belong to the same cluster in $\mathcal{C}$ and to different groups in $\mathcal{P}$, and (d) *DS* if the vectors belong to different clusters in $\mathcal{C}$ and to the same group in $\mathcal{P}$. Let a, b, c, and d be the number of *SS*, *SD*, *DS*, and *DD* pairs of vectors of X, respectively. Then $a + b + c + d = M$, where M is the total number of possible pairs in X, that is, $M = N(N-1)/2$.

Example 16.1

Let $X = \{\boldsymbol{x}_i, i = 1, \ldots, 6\}$, $\mathcal{C} = \{\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}, \{\boldsymbol{x}_4, \boldsymbol{x}_5\}, \{\boldsymbol{x}_6\}\}$, and $\mathcal{P} = \{\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}, \{\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}\}$. The following table shows the type of all pairs of vectors in X.

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$
$\boldsymbol{x}_1$		SS	SS	DD	DD	DD
$\boldsymbol{x}_2$			SS	DD	DD	DD
$\boldsymbol{x}_3$				DD	DD	DD
$\boldsymbol{x}_4$					SS	DS
$\boldsymbol{x}_5$						DS
$\boldsymbol{x}_6$						

From this table we obtain $a = 4$, $b = 0$, $c = 2$, and $d = 9$.

Let $m_1 = a + b$ be the number of pairs of vectors that belong to the same cluster in $\mathcal{C}$ and $m_2 = a + c$ be the number of pairs of vectors that belong to the same group in $\mathcal{P}$. Using the preceding definitions, we can define statistical indices (statistics) in order to measure the degree to which $\mathcal{C}$ matches $\mathcal{P}$. Such statistical indices are the following:

- Rand statistic

$$R = (a + d)/M \tag{16.5}$$

- Jaccard coefficient

$$J = a/(a + b + c) \tag{16.6}$$

- Fowlkes and Mallows index

$$FM = a/\sqrt{m_1 m_2} = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} \tag{16.7}$$

The term $a+d$ is the number of SS pairs of vectors plus the number of DD pairs. Thus, the Rand statistic measures the fraction of the total number of pairs that are either SS or DD. The Jaccard coefficient follows the same philosophy as the Rand statistic, except that it excludes d. The values of these two statistics are between 0 and 1. However, a prerequisite for achieving the maximum value is to have $m = s$, which, in general, is not always the case.

For all the above defined indices, it is clear that the larger their value, the higher the agreement between $\mathcal{C}$ and $\mathcal{P}$, that is, *all the corresponding statistical tests are right tailed*.

Another very popular statistic that is, frequently used in conjunction with external criteria is Hubert's Γ statistic (e.g., [Hube 76, Mant 67, Bart 62]). It measures the correlation between two matrices, X and Y, of dimension $N \times N$, drawn independently of each other. For symmetric matrices this can be written as

- Hubert's Γ statistic

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j)Y(i,j) \tag{16.8}$$

where $X(i,j)$ and $Y(i,j)$ are the (i,j) elements of the matrices X and Y, respectively. High values of Γ indicate close agreement between X and Y. The normalized version of the Γ statistic, denoted by $\hat{\Gamma}$, is also used.

- Normalized Γ statistic

$$\hat{\Gamma} = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X(i,j)-\mu_X)(Y(i,j)-\mu_Y)}{\sigma_X \sigma_Y} \tag{16.9}$$

where μ_X, μ_Y, σ_X^2, and σ_Y^2 are the respective means and variances, that is, $\mu_X = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} X(i,j)$, $\sigma_X^2 = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} X(i,j)^2 - \mu_X^2$ (similarly we define μ_Y and σ_Y^2). The values of $\hat{\Gamma}$ are between -1 and 1.

Let us set $X(i,j)$ equal to 1 if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster in $\mathcal{C}$ and 0 otherwise, and $Y(i,j)$ equal to 1 if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same group in $\mathcal{P}$ and 0 otherwise. It can then be shown (see Problem 16.2) that in this case the $\hat{\Gamma}$ statistic becomes equal to

$$\hat{\Gamma} = (Ma - m_1 m_2)/\sqrt{m_1 m_2 (M - m_1)(M - m_2)} \tag{16.10}$$

Unusually large absolute values of $\Gamma(\hat{\Gamma})$ suggest that $\mathcal{C}$ and $\mathcal{P}$ agree with each other.

As almost always happens in practice, the exact computation of the pdf of all these indices, under the null hypothesis, is very difficult. Thus, we use Monte Carlo techniques for their estimation. In the sequel, we discuss such a procedure, which is based on the random position hypothesis. Data are assumed to be ratio scaled.

- For $i = 1$ to r
  - Generate a data set X_i of N vectors in the area of interest of X, so that the vectors are uniformly distributed in it.
  - Assign each vector $\boldsymbol{y}_j^i \in X_i$ to the group where the $\boldsymbol{x}_j \in X$ belongs, according to the structure imposed by $\mathcal{P}$.
  - Run the same clustering algorithm, used for obtaining $\mathcal{C}$, on X_i and let $\mathcal{C}_i$ be the resulting clustering.
  - Compute the value $q(\mathcal{C}_i)$ of the corresponding statistical index q for $\mathcal{P}$ and $\mathcal{C}_i$.
- End {For}
- Create the histogram of $q(\mathcal{C}_i)$'s.

The following example demonstrates how this methodology can be used in practice.

Example 16.2

(a) Consider a data set X of 100 vectors in the H_3 hypercube. The data are generated to form four groups, each consisted of 25 vectors. Each group is generated by a normal distribution. The first group of 25 vectors of X is generated from the first distribution while the second, third, and fourth groups of 25 vectors are generated from the second, the third, and the fourth distribution, respectively. The covariance matrices of all distributions are equal to $0.2I$, where I is the 3×3 identity matrix. The mean vectors for the four distributions are $[0.2, 0.2, 0.2]^T$, $[0.5, 0.2, 0.8]^T$, $[0.5, 0.8, 0.2]^T$, and $[0.8, 0.8, 0.8]^T$, respectively. If a distribution generates a vector that is, outside the unit hypercube, it is ignored and replaced by another that lies inside H_3. It is not difficult to realize that the points of X form four compact and well-separated clusters.

We assume that the external information is: "The vectors of X belong to four different groups P_1, P_2, P_3, and P_4, such that P_1 contains the first 25 vectors of X and P_2, P_3, and P_4 contain the second, third, and fourth groups of 25 vectors of X, respectively."

We run the isodata algorithm for $m = 4$ and let $\mathcal{C}$ be the resulting clustering. We compute the values of the Rand, R, the Jaccard, J, the Fowlkes and Mallows, FM, and the $\hat{\Gamma}$ statistics for $\mathcal{C}$ and $\mathcal{P}$. These are 0.91, 0.68, 0.81, and 0.75, respectively. Next, we estimate the distribution of these statistics using the procedure described before. Specifically, 100 data sets X_i, $i = 1, \ldots, 100$, are generated, each of them consisting of 100 randomly selected vectors in H_3, following the uniform distribution. According to the $\mathcal{P}$ defined earlier, we assign the first 25 of them to P_1 and the second, third, and fourth groups of 25 vectors to P_2, P_3, and P_4, respectively. For each X_i we run the isodata algorithm for $m = 4$ and we produce the clustering $\mathcal{C}_i$, $i = 1, \ldots, 100$. Then we compute the values of the four statistics, R_i, J_i, FM_i, and $\hat{\Gamma}_i$ for each $\mathcal{C}_i$ and $\mathcal{P}$, $i = 1, \ldots, 100$. We set the significance level at $\rho = 0.05$.

Then, in terms of a given statistic, we accept or reject the null hypothesis (i.e., the random hypothesis) according to the conditions given in Section 16.2. In our case R is greater than all R_i's. Similarly, J, FM, and $\hat{\Gamma}$ are greater than all J_i's, FM_i's, and $\hat{\Gamma}_i$'s, respectively. Thus, all statistics reject the null hypothesis at significance level $\rho = 0.05$.

(b) Now let X' be a data set constructed as X, but with the covariance matrices of the normal distributions equal to $0.6I$. In this case, the vectors of X form weak clusters, that is, clusters that exhibit "large" spread around their mean vector. The values of the four statistics in this case are $R = 0.64$, $J = 0.15$, $FM = 0.27$, and $\hat{\Gamma} = 0.03$. R is greater than 99 of the R_i's. Similarly, J, FM, and $\hat{\Gamma}$ are greater than 94 J_i's, 94 FM_i's, and 98 $\hat{\Gamma}_i$'s, respectively. Thus, according to the Rand and $\hat{\Gamma}$ statistics, the null hypothesis is rejected at significance level $\rho = 0.05$. However, this is not the case for the other two indices.

This situation illustrates the fact that different statistics may lead to different conclusions when no clear-cut situations are considered (see also comparative studies in [Mill 80, Mill 83, Mill 85]).

(c) Let us now construct X'' by selecting the covariance matrices equal to $0.8I$. In this case, the vectors of X'' are so dispersed that, practically, X'' does not exhibit any clustering structure. The values of the four statistics in this case are $R = 0.63$, $J = 0.14$, $FM = 0.25$, and $\hat{\Gamma} = -0.01$. Specifically, R is greater than 62, from the total of 100, R_i's. Similarly, J, FM, and $\hat{\Gamma}$ are greater than 48 R_i's, 48 J_i's, and 55 $\hat{\Gamma}_i$'s, respectively. Thus, according to all statistics, the null hypothesis is not rejected at significance level $\rho = 0.05$.

Remark

- For each of these statistics, q, there exists a corresponding "corrected" statistic q', which is a normalized version of q and is defined as

$$q' = \frac{q - E(q)}{\max(q) - E(q)} \tag{16.11}$$

where $\max(q)$ is the maximum possible value of q and $E(q)$ is the mean value of q, under the null hypothesis. Its values are between 0 and 1. The maximum value is always achievable when a perfect match between $\mathcal{C}$ and $\mathcal{P}$ occurs and the minimum if $\mathcal{C}$ and $\mathcal{P}$ have been chosen by chance. The problem encountered here is the computation of $E(q)$ and $\max(q)$. This problem is attacked in [Hube 85], for the Rand statistic, under the assumption that the maximum value of the Rand statistic is 1. The same problem for the Fowlkes–Mallows index is treated in [Fowl 83].

Assessing the Agreement between $\mathcal{P}$ and Proximity Matrix P

In this section, we show that the Γ statistic can be used to measure the degree to which the proximity matrix P of X matches a partition $\mathcal{P}$, which is imposed *a priori* on X. Recall that $\mathcal{P}$ may be viewed as a mapping g of X to $\{1, \ldots, m\}$. Let us consider the matrix Y whose (i,j) element, $Y(i,j)$, is defined as follows:

$$Y(i,j) = \begin{cases} 1, & \text{if } g(\boldsymbol{x}_i) \neq g(\boldsymbol{x}_j) \\ 0, & \text{otherwise} \end{cases} \tag{16.12}$$

for $i, j = 1, \ldots, N$. It is clear that Y is symmetric. Then, the Γ (or $\hat{\Gamma}$) statistic is applied to the proximity matrix P and Y. Its value is a measure of the degree to which Y matches P.

In order to estimate the pdf of Γ (or $\hat{\Gamma}$) under the *random label hypothesis*, we produce, say, r mappings $g_i, i = 1, \ldots, r$.[3] For each of them we form the corresponding Y_i matrix and we apply the Γ (or $\hat{\Gamma}$) statistic to P and Y_i, $i = 1, \ldots, r$. Then we proceed as usual for the acceptance or rejection of the random label hypothesis.

Example 16.3

We consider a data set X of 64 two-dimensional vectors. The first 16 of them spring out of a normal distribution with mean $[0.2,\ 0.2]^T$, and the remaining three groups of 16 vectors stem from three normal distributions with means $[0.2, 0.8]^T$, $[0.8, 0.2]^T$, and $[0.8, 0.8]^T$, respectively. The covariance matrices of all distributions are equal to $0.15I$. Let P be the proximity matrix of X when the squared Euclidean distance is in use. Also, we set the significance level at $\rho = 0.05$.

(a) Let $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$. Suppose that the first set of 16 vectors is assigned to P_1, the second is assigned to P_2, the third is assigned to P_3, and the last to P_4. Based on this information, we form Y as described before and we compute the value of $\hat{\Gamma}$ for P and Y, which is found to be 0.77. Then we generate random partitions $\mathcal{P}_i$, $i = 1, \ldots, 100$, we form the corresponding matrices Y_i, and we compute the values $\hat{\Gamma}_i$ between P and each of the Y_i's. It turns out that $\hat{\Gamma}$ is greater than all of these values. Thus, the null hypothesis is rejected at significance level ρ.

(b) Assume now that the external information $\mathcal{P}$ assigns randomly 16 vectors of X to each P_i. It is clear that the external information does not agree with the underlying structure of X. If we apply the same procedure as before, we find that $\hat{\Gamma} = -0.01$, which is less than 70 values of $\hat{\Gamma}_i$. Thus, the randomness hypothesis is accepted.

16.3.2 Internal Criteria

Our aim here is to verify whether the clustering structure produced by a clustering algorithm fits the data, using only information inherent in the data. In the sequel, unless otherwise stated, we consider the case in which the data are represented by their proximity matrix. Two cases are considered: (a) the clustering structure is a hierarchy of clusterings and (b) the clustering structure consists of a single clustering.

Validation of Hierarchies of Clusterings

We recall that the dendrogram produced by a hierarchical clustering algorithm may be represented by the respective cophenetic matrix, P_c. *We will define statistical indices that measure the degree of agreement between the cophenetic matrix,*

[3] Typically, $r = 100$.

P_c, produced by a specific hierarchical clustering algorithm, with the proximity matrix P of X. Because both matrices are symmetric and have their diagonal elements equal to 0,[4] we consider only the $M \equiv N(N-1)/2$ upper diagonal elements of P_c and P. Let d_{ij} and c_{ij} be the (i,j) element of P and P_c, respectively.

The first index, known as the *cophenetic correlation coefficient* (*CPCC*) measures the correlation between P_c and P and is used when the matrices are interval or ratio scaled. It is defined as

$$CPCC = \frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}c_{ij} - \mu_P\mu_c}{\sqrt{\left((1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_P^2\right)\left((1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_c^2\right)}} \tag{16.13}$$

where the corresponding mean values are defined as in Eq. (16.9). It can be shown that the values of the *CPCC* are between -1 and 1 (see Problem 16.4). The closer the *CPCC* index to 1, the better the agreement between the cophenetic and the proximity matrix. The *CPCC* statistic has been studied by various researchers (see, e.g., [Rolp 68, Rolp 70, Farr 69]). The major difficulty associated with it is that it depends on many parameters of the problem, such as the size of X, the clustering algorithm used and the employed proximity measure. Hence, the exact computation of its pdf under H_0 is very difficult. Once more, one is forced to use Monte Carlo techniques for the estimation of its distribution, under H_0. According to the random position hypothesis, we generate r sets X_i, whose vectors are randomly distributed according to the uniform distribution, and we apply to each X_i the same hierarchical algorithm that has produced P_c. Then, we compute *CPCC* for the proximity matrix of X_i, P_i, and the resulting cophenetic matrix, P_{c_i} and we construct the corresponding histogram.

Interestingly enough, in [Rolp 70], it is stated that even high values of *CPCC* (near 0.9) should be handled with caution when the unweighted pair group method average (UPGMA) algorithm is in use (Chapter 13), as there are cases for which even such large values cannot guarantee close agreement between the cophenetic and the proximity matrix.

Another statistical index, which is suitable for cases in which P_c and P are ordinally scaled, is the γ statistic, which is described in the sequel. Let $\boldsymbol{v}_P$ and $\boldsymbol{v}_c$ be two vectors of dimension $N(N-1)/2$, each containing the upper diagonal elements of P and P_c, respectively, ordered by rows. Let (v_{P_i}, v_{P_j}) and (v_{c_i}, v_{c_j}) be two pairs of elements of $\boldsymbol{v}_P$ and $\boldsymbol{v}_c$, respectively. The following definitions are in order.

A set of pairs $\{(v_{P_i}, v_{P_j}), (v_{c_i}, v_{c_j})\}$ is called

- concordant if

$$\left((v_{P_i} < v_{c_i})\ \&\ (v_{P_j} < v_{c_j})\right)\ or\ \left((v_{P_i} > v_{c_i})\ \&\ (v_{P_j} > v_{c_j})\right)$$

[4] This implies that we use a dissimilarity measure.

- discordant if

$$\big((v_{p_i} < v_{c_i})\ \&\ (v_{p_j} > v_{c_j})\big)\ or\ \big((v_{p_i} > v_{c_i})\ \&\ (v_{p_j} < v_{c_j})\big)$$

Finally, a set of pairs is neither concordant nor discordant if $v_{p_i} = v_{c_i}$ or $v_{p_j} = v_{c_j}$. Let S_+ and S_- be the numbers of the concordant and discordant pairs, respectively. Then γ is defined as

$$\gamma = \frac{S_+ - S_-}{S_+ + S_-} \tag{16.14}$$

The γ statistic takes values between -1 and 1.

Example 16.4

Let $\boldsymbol{v}_p = [3, 2, 1, 5, 2, 6]^T$ and $\boldsymbol{v}_c = [2, 3, 5, 1, 6, 4]^T$. For all possible 16 pairs of pairs we have

Index	$\boldsymbol{v}_p$	$\boldsymbol{v}_c$		Index	$\boldsymbol{v}_p$	$\boldsymbol{v}_c$	
(1, 2)	(3, 2)	(2, 3)	dis.	(2, 6)	(2, 6)	(3, 4)	dis.
(1, 3)	(3, 1)	(2, 5)	dis.	(3, 4)	(1, 5)	(5, 1)	dis.
(1, 4)	(3, 5)	(2, 1)	con.	(3, 5)	(1, 2)	(5, 6)	con.
(1, 5)	(3, 2)	(2, 6)	dis.	(3, 6)	(1, 6)	(5, 4)	dis.
(1, 6)	(3, 6)	(2, 4)	con.	(4, 5)	(5, 2)	(1, 6)	dis.
(2, 3)	(2, 1)	(3, 5)	con.	(4, 6)	(5, 6)	(1, 4)	con.
(2, 4)	(2, 5)	(3, 1)	dis.	(5, 6)	(2, 6)	(6, 4)	dis.
(2, 5)	(2, 2)	(3, 6)	con.				

Thus, $S_+ = 6$, $S_- = 9$ and $\gamma = -1/5 = -0.2$.

The γ statistic depends on all the factors of the problem at hand and, as a consequence, the estimate of its pdf under the randomness hypothesis (H_0) is also difficult to derive. Thus, one has to use Monte Carlo techniques once again for the estimation of the pdf of γ under H_0. In this case, the random graph hypothesis is used. Specifically, we produce r random rank order proximity matrices P_i, with no ties, and we run the algorithm that produced P_c on each of them. Then we compute the value of γ for each P_i and its corresponding cophenetic matrix P_{c_i} and we form the histogram for the values of γ.

Remarks

- It has been conjectured [Hube 74] that when the single and the complete link algorithms are used, the statistic $N\gamma - a \ln N$ follows (approximately) the standard normal distribution. The constant a is set equal to 1.1 (1.8) when the single (complete) link algorithm is used. If we adopt this conjecture, it relieves us of the computational burden of the Monte Carlo method.

- The γ statistic may also be used to compare the results for two different hierarchies of clusterings resulting from two different clustering algorithms. (e.g., [Bake 74, Hube 74], Problem 16.5).

Another measure that is, suitable for ordinal-scaled P and P_c is Kudall's τ statistic [Cunn 72], which is defined as

$$\tau = \frac{S_+ - S_-}{N(N-1)/2} \tag{16.15}$$

The difference from the γ statistic is that the denominator here extends to all sets of pairs, whereas in the case of the γ statistic the sets of pairs that are neither concordant nor discordant are excluded.

Validation of Individual Clusterings

Our goal here is to investigate whether a given clustering $\mathcal{C}$, consisting of m clusters, matches information that is, inherent in the data set X. In the sequel, we show that the Γ (or $\hat{\Gamma}$) statistic can be used in order to achieve this goal. Once again, we use the proximity matrix P as a measure representing the structural information inherent in the data. The (i,j) element of the matrix Y is defined as

$$Y(i,j) = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ belong to different clusters} \\ 0, & \text{otherwise} \end{cases} \tag{16.16}$$

for $i,j = 1,\ldots,N$. It is clear that Y is symmetric. Then the Γ (or $\hat{\Gamma}$) statistic is applied to P and Y. Its value is a measure of the degree of correspondence between P and Y.

The *random position hypothesis* is employed. For each of the resulting random data sets X_i, the proximity matrix P_i is computed. Then we apply, to each of them, the clustering algorithm used to produce $\mathcal{C}$. Let $\mathcal{C}_i$, $i = 1,\ldots,r$, be the resulting clusterings of m clusters. We compute Y_i and Γ_i. Finally, we decide for the rejection or acceptance of the null hypothesis at a given significance level ρ according to the conditions given in Section 16.2.

Example 16.5

Consider a data set X of 100 vectors in the H_2 hypercube. The vectors are generated to form four groups, each of 25 vectors. Each group is generated by a normal distribution. The corresponding covariance matrices are all equal to $0.1I$ and the mean vectors are $[0.2,\ 0.2]^T$, $[0.8, 0.2]^T$, $[0.2, 0.8]^T$, $[0.8, 0.8]^T$, respectively. We apply the isodata algorithm and let $\mathcal{C}$ be the resulting clustering. Computing the corresponding matrices Y and P, we obtain $\hat{\Gamma} = 0.5704$. Then we generate 100 data sets, X_i, whose vectors are randomly distributed in H_2, following the uniform distribution. The isodata algorithm is applied to each of them, and let $\mathcal{C}_i$, $i = 1,\ldots,100$, be the resulting clusterings. Computing Y_i and P_i associated with the resulting clusterings for each X_i, it turns out that 99 of the corresponding $\hat{\Gamma}_i$ values are smaller than $\hat{\Gamma}$. Thus, the null hypothesis is rejected at significance level $\rho = 0.05$.

Repeating the experiment but with covariance matrices equal to $0.2I$, we find that $\hat{\Gamma}$ is greater than 86 of 100 $\hat{\Gamma}_i$ values. Thus, the null hypothesis is not rejected at significance level $\rho = 0.05$.

16.4 RELATIVE CRITERIA

So far, clustering validation has been performed on the basis of statistical tests. A major drawback of most of these techniques is their high computational demands, due to the required Monte Carlo methodology. In this section, a different approach is discussed that does not involve statistical tests. To this end, *a set of clusterings is considered and the goal is to choose the best one according to a prespecified criterion*. More specifically, let $\mathcal{A}$ be the set of parameters associated with a specific algorithm. For example, for the algorithms of Chapter 14, $\mathcal{A}$ contains the number of clusters, m, as well as the initial estimates of the parameter vectors associated with each cluster. The problem can be stated as follows:

"Among the clusterings produced by a specific clustering algorithm, for different values of the parameters in $\mathcal{A}$, choose the one that best fits the data set X."

We consider the following cases:

- *$\mathcal{A}$ does not contain the number of clusters*, m, as a parameter (such as the algorithms based on graph theory, the morphological clustering algorithm and the boundary detection algorithms).

  The choice of the "best" parameter values for this type of algorithm is based on the assumption that *if X possesses a clustering structure, this structure is captured for a "wide" range of values of the parameters* in $\mathcal{A}$ (e.g., [Post 93]). Based on this assumption, we proceed as follows. We run the algorithm for a wide range of values of its parameters and we choose the widest range for which, m, remains constant (typically $m \ll N$). Then we choose as appropriate values of the parameters of $\mathcal{A}$ the values that correspond to the middle of this range. Note that, implicitly, this procedure also identifies the number of clusters that underlie X.

Example 16.6

(a) We consider a data set X, consisted of three groups of 100 two-dimensional vectors. These groups are formed from normal distributions with means $[0, 0]^T$, $[8, 4]^T$, and $[8, 0]^T$, respectively, and covariance matrices equal to $1.5I$. As one can easily observe in Figure 16.2a, the three groups form three compact and well-separated clusters. We run the binary morphology clustering algorithm (BMCA), using the 3×3 structuring element (Figure 15.10a), with the resolution parameter r ranging from 1 to 77 and we plot the number of clusters versus r (Figure 16.2b). We observe that for any value of r between 37 and 67, the number of clusters remains constant and equal to 3. Taking into account that this range of values is the largest one, we choose $r = 52$, and we conclude that our data form three clusters.

(b) Generate another data set, as before, but with the covariance matrices equal to $2.5I$. This data set is depicted in Figure 16.3a. We observe that in this case the three groups are so dispersed that they practically cannot be distinguished from each other. We run BMCA once again, using the 3×3 structuring element, for r ranging from 1 to 77, with step 1, and we

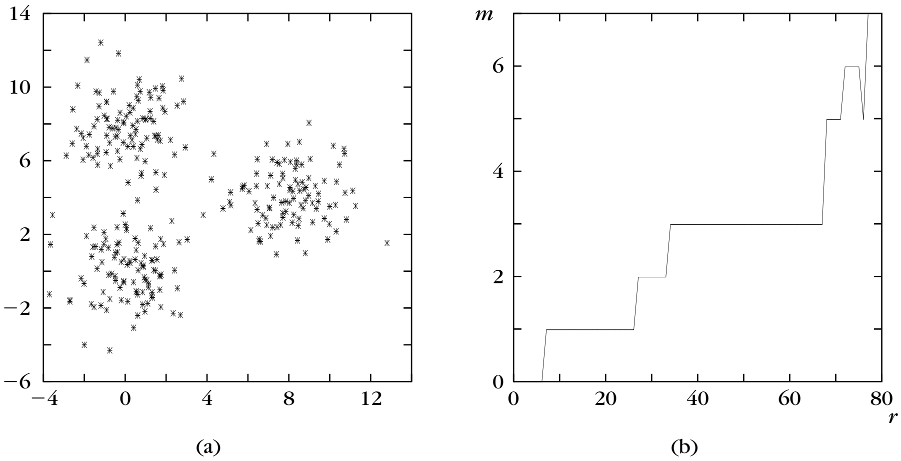


FIGURE 16.2

(a) Three well-separated clusters. (b) The plot of the number of clusters m versus the resolution parameter r, using the binary morphology clustering algorithm (BMCA).

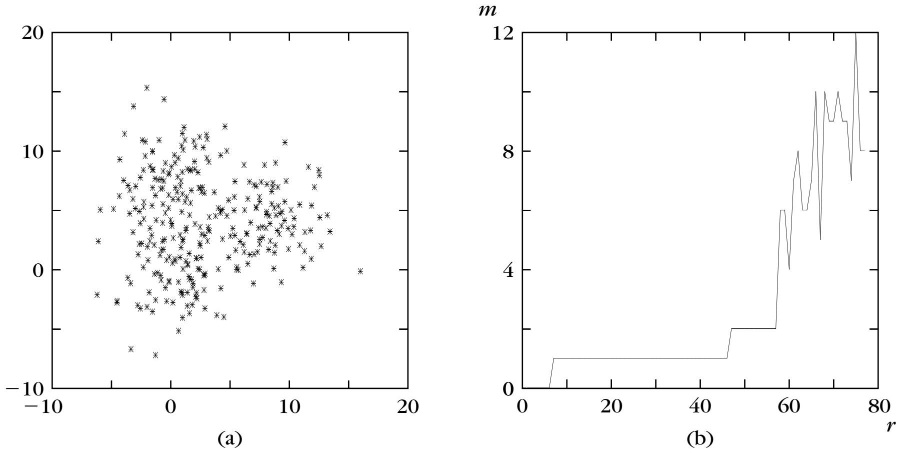


FIGURE 16.3

(a) Three overlapped clusters. (b) The plot of m versus r.

plot the number of clusters versus r (see Figure 16.3b). In this case, for $r = 7, \ldots, 46$, the number of clusters remains constant and the corresponding value of m is 1.

- *$\mathcal{A}$ contains m as a parameter* (such as the fuzzy and hard clustering algorithms discussed in Chapter 14). For this case, a different procedure is followed. We first select a suitable performance index q. The "best" clustering is identified, in terms of q, via the following procedure. We run the clustering algorithm at hand for all values of m between a minimum $m_{\min}$ and a maximum $m_{\max}$, where $m_{\min}$ and $m_{\max}$ are chosen *a priori*. For each value of m, we run the algorithm r times, using different sets of values for the other parameters of $\mathcal{A}$.[5] Then we plot the best values of q, obtained for each m, versus m and we seek the maximum or the minimum of this plot, according to whether large or small values of q indicate good clusterings. This procedure works well if q exhibits no trend with respect to m. However, as we will see, several of the commonly used indices q exhibit an increasing (decreasing) trend as m increases. Thus, locating the maximum (minimum) versus m is no longer indicative of a good clustering. For indices that exhibit such behavior, in the range $[m_{\min}, m_{\max}]$, we search for values of m at which a significant local change in the value of q occurs. This change appears in the plot as a significant "knee." *The presence of such a knee is an indication of the number of clusters underlying X. On the other hand, the absence of such a knee may be an indication that X possesses no clustering structure.*

Another source of complication, associated with many of the indices used in this framework is that their behavior depends on many other factors such as the number of vectors in X and their dimensionality. The situation is demonstrated via the following example.

Example 16.7

(a) In this example, we consider 16 different data sets with different numbers of vectors and dimensionalities. Specifically, we consider four 2-dimensional data sets of $50, 100, 150$, and 200 vectors; four 4-dimensional data sets of $50, 100,\ 150$, and 200 vectors; four 6-dimensional data sets of $50, 100, 150$, and 200 vectors; and four 8-dimensional data sets of $50, 100, 150$, and 200 vectors. The vectors of the data sets lie in the H_i hypercube, $i = 2, 4, 6, 8$, respectively. All the data sets contain four compact and well-separated clusters. All the clusters stem from normal distributions with means $[\overbrace{0.2,\ldots,0.2}^{i}]^T$, $[\overbrace{0.2,\ldots,0.2}^{i/2}, \overbrace{0.8,\ldots,0.8}^{i/2}]^T$, $[\overbrace{0.8,\ldots,0.8}^{i/2}, \overbrace{0.2,\ldots,0.2}^{i/2}]^T$, $[\overbrace{0.8,\ldots,0.8}^{i}]^T$, where i is the dimensionality, and covariance matrices $0.2I_i$, where I_i is the $i \times i$ identity matrix. For each of these data sets we run the isodata algorithm for $m = 1, \ldots, 10$, and we compute the corresponding values of the cost function J. For this case, only a single run is performed for each m. In Figure 16.4 we plot J versus the number of clusters, m, for the cases of 50, 100, 150, and 200 vectors and for different dimensionalities.

One can easily notice that the higher the dimensionality, the sharper the knee at $m = 4$. Moreover, as the size of the data set increases, the knee at $m = 4$ becomes sharper, even at lower dimensionalities. Rules for automatic identification of a knee are discussed in [Dube 87a].

[5] For example, if the k-means algorithm is used, we run it using different initial conditions.

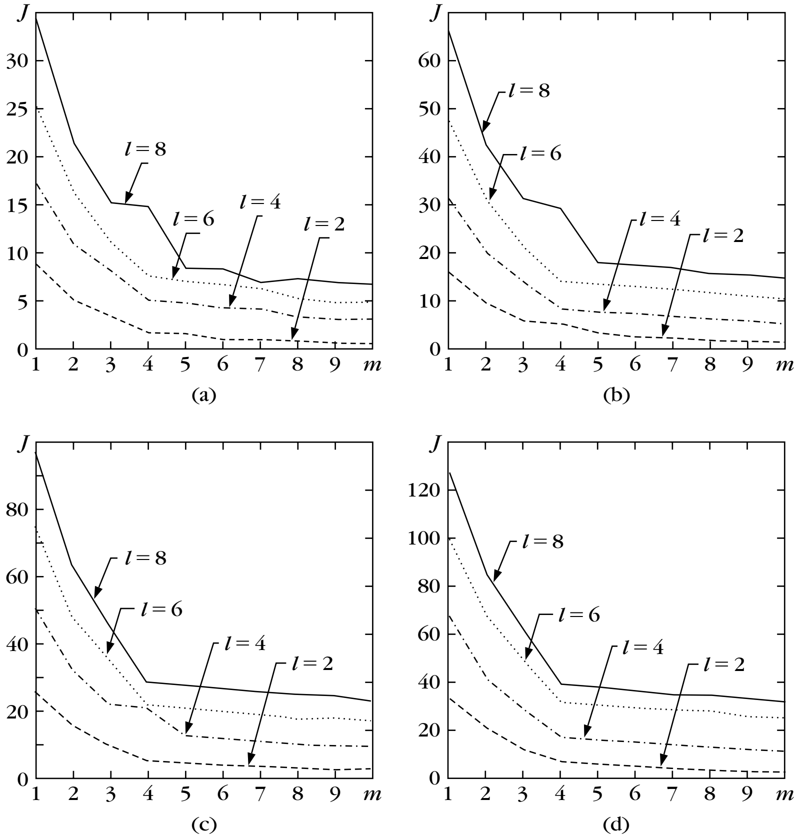


FIGURE 16.4

Plots of J versus m for (a) $N = 50$, (b) $N = 100$, (c) $N = 150$, (d) $N = 200$, for clustered data.

(b) We again construct 16 data sets, but now the vectors in each of them are randomly distributed in the H_i hypercube, according to the uniform distribution. If we carry out the same procedure as before, we see in Figure 16.5 that there are no sharp knees in the plots. Thus, the absence of sharp knees in the plots *may be* an indication of the absence of clustering structure.

16.4.1 Hard Clustering

In this section we discuss indices that are suitable for hard clusterings. In the sequel unless otherwise stated, we consider only the case of compact clusters.

- *The modified Hubert* Γ *statistic.* Let $c_i = k$ if the vector $\boldsymbol{x}_i$ belongs to cluster C_k. Also let Q be the $N \times N$ matrix whose (i,j) element, $Q(i,j)$, is equal to the distance $d(\boldsymbol{w}_{c_i}, \boldsymbol{w}_{c_j})$ between the representatives of the clusters where $\boldsymbol{x}_i$

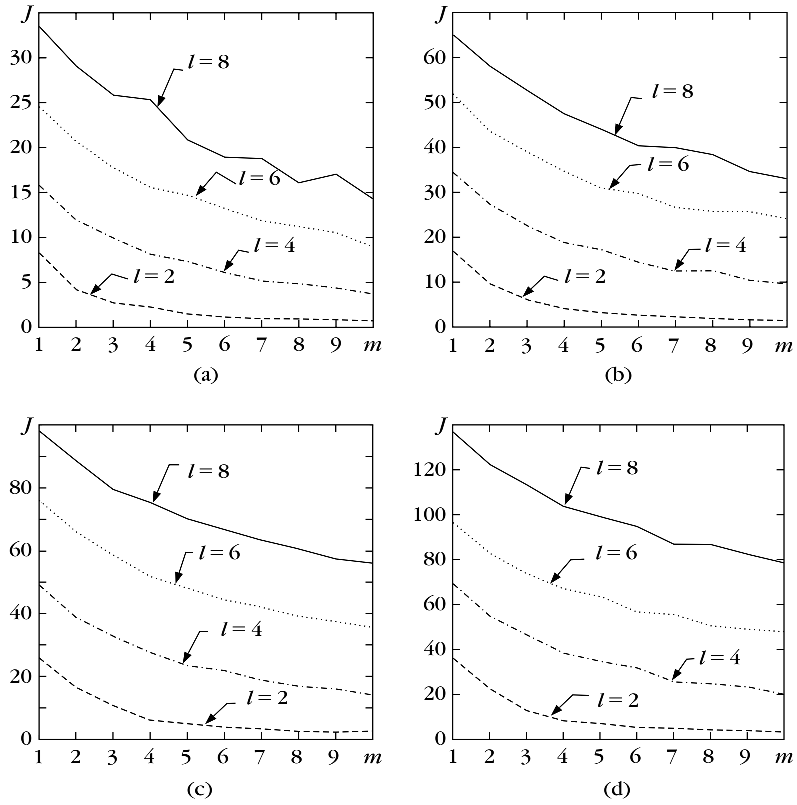


FIGURE 16.5

Plots of J versus m for (a) $N = 50$, (b) $N = 100$, (c) $N = 150$, (d) $N = 200$, for random data.

and $\boldsymbol{x}_j$ belong. The modified Hubert Γ statistic is defined as in Eq. (16.8) and it is applied to the proximity matrix P of the data set X and the matrix Q (of course, the same distance measure must be used for both P and Q). Similarly, we can define the normalized modified Hubert Γ statistic. It is clear that if $d(\boldsymbol{w}_{c_i}, \boldsymbol{w}_{c_j})$ is close to $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$, for $i,j = 1, \ldots, N$, that is, when compact clusters are encountered in X, P and Q will be in close agreement and the values of Γ and $\hat{\Gamma}$ will be high. Conversely, high values of $\Gamma(\hat{\Gamma})$ indicate the existence of compact clusters. If the opposite is true, the values of the modified Γ and $\hat{\Gamma}$ indices are expected to be low. Thus, in the plot of $\hat{\Gamma}$ versus m, we seek a significant knee that corresponds to a significant increase of $\hat{\Gamma}$. The value of m at which this knee occurs indicates the number of clusters tha underlie X.

For $m = 1$ and $m = N$ the index is not defined. Also, this index tends to increase as m increases toward N (see Problem 16.6) for random data and tends to be flat for data sets that possess a clustering structure [Jain 88].

- *The Dunn and Dunn-like indices.* Let the dissimilarity function between two clusters C_i and C_j be (Chapter 11)

$$d(C_i, C_j) = \min_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y}) \tag{16.17}$$

and define the *diameter of a cluster* C as

$$diam(C) = \max_{\boldsymbol{x}, \boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y}) \tag{16.18}$$

that is, the diameter of a cluster C is the distance of its two most distant vectors. $diam(C)$ may be viewed as a measure of dispersion of C. Then, the Dunn index for a specific m is defined as

$$D_m = \min_{i=1,\ldots,m} \left\{ \min_{j=i+1,\ldots,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,\ldots,m} diam(C_k)} \right) \right\} \tag{16.19}$$

It is clear that if X contains compact and well-separated clusters, Dunn's index will be large, since the distance between the clusters is expected to be "large" and the diameter of the clusters is expected to be "small." Conversely, large values of Dunn's index indicate the presence of compact and well-separated clusters. The index D_m does not exhibit any trend with respect to m, hence the maximum in the plot of D_m versus m can be used to indicate the number of clusters that underlie X.

In [Dunn 74], it is shown that if $D_m > 1$ for a specific clustering, then this clustering contains compact and well-separated clusters.

A disadvantage of the Dunn index is the considerable amount of time required for its computation (see Problem 16.7). Moreover, Dunn's index is sensitive to the presence of noisy vectors in X, because these are likely to increase the value of the denominator of Eq. (16.19).

In [Pal 97] three Dunn-like indices are proposed that are more robust to the presence of noisy vectors. Furthermore, preliminary simulation results show that they may be used for cases in which shell-shaped clusters underlie X. These three indices are based on the concepts of the minimum spanning tree (MST), the relative neighborhood graph (RNG), and the Gabriel graph (GG), discussed in Chapter 15. Let us consider explicitly the index based on the MST concept. The other two are defined using similar arguments.

Consider a cluster C_i and the *complete* graph G_i having vertices that correspond to the vectors of C_i. The weight, w_e, of an edge, e, of this graph equals the distance between its two end points, $\boldsymbol{x}$ and $\boldsymbol{y}$, that is, $w_e = d(\boldsymbol{x}, \boldsymbol{y})$. Let E_i^{MST} be the set of edges of the MST of G_i and let e_i^{MST} be the edge in E_i^{MST} with the maximum weight. Then the diameter of C_i, $\text{diam}_i^{\text{MST}}$, is defined as the weight of e_i^{MST} (see Figure 16.6).

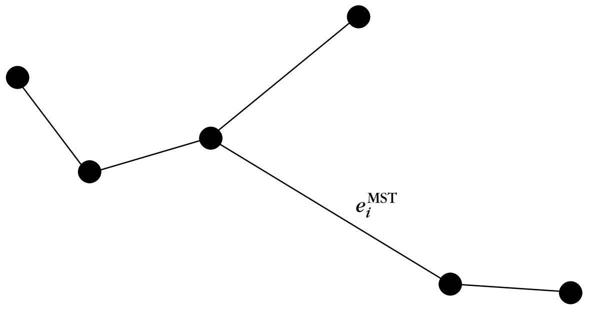


FIGURE 16.6

A minimum spanning tree.

The dissimilarity between two clusters is defined as the distance of their mean vectors, $d(C_i, C_j) = d(\boldsymbol{m}_i, \boldsymbol{m}_j)$. Then, the Dunn-like index, based on the concept of the MST, is defined as

$$D_m^{\mathrm{MST}} = \min_{i=1,\ldots,m} \left\{ \min_{j=i+1,\ldots,m} \left(\frac{d(C_i, C_j)}{\max_{k=1,\ldots,m} diam_k^{\mathrm{MST}}} \right) \right\} \tag{16.20}$$

The maximum in the plot of D_m^{MST} versus m indicates the underlying number of clusters in X. Similar arguments are followed to define Dunn-like indices for GG and RNG graphs (see Problem 16.8).

- *The Davies-Bouldin (DB) and DB-like indices.* Let s_i be a measure of dispersion of a cluster C_i (i.e., a measure of its spread around its mean vector) and $d(C_i, C_j) \equiv d_{ij}$ the dissimilarity between two clusters, using an appropriate dissimilarity measure. Based on these, a similarity index R_{ij} between C_i and C_j is defined to satisfy the following conditions [Davi 79]:

  (C1) $R_{ij} \geq 0$.
  (C2) $R_{ij} = R_{ji}$.
  (C3) If $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$.
  (C4) If $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$.
  (C5) If $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$.

  These conditions state that R_{ij} is nonnegative and symmetric. If both clusters, C_i and C_j, collapse to a single point, then $R_{ij} = 0$. A cluster C_i with the same distance from two other clusters, C_j, C_k, is more similar to the cluster with the largest dispersion (condition (C4)). For the case of equal dispersions and different dissimilarity levels, the cluster C_i is more similar to the closer of the two (condition (C5)).

A (simple) choice for an R_{ij} that satisfies these conditions is the following [Davi 79]:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{16.21}$$

provided that d_{ij} is symmetric.

Also let R_i be defined as

$$R_i = \max_{j=1,\ldots,m,j \neq i} R_{ij}, \quad i = 1, \ldots, m \tag{16.22}$$

Then the DB index is defined as

$$DB_m = \frac{1}{m} \sum_{i=1}^{m} R_i \tag{16.23}$$

That is, DB_m is the average similarity between each cluster $C_i, i = 1, \ldots, m$, and its most similar one. As it is desirable for the clusters to have the minimum possible similarity to each other, we seek clusterings that minimize DB. On the other hand, small values of DB are indicative of the presence of compact and well-separated clusters. The DB_m index exhibits no trends with respect to m [Davi 79], thus we seek the minimum value of DB_m, in the plot of DB_m versus m.

In [Davi 79], the dissimilarity $d(C_i, C_j)$ between two clusters is defined as

$$d_{ij} = \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_q = \left(\sum_{k=1}^{l} |w_{ik} - w_{jk}|^q \right)^{1/q} \tag{16.24}$$

Also, the dispersion of a cluster C_i is defined as

$$s_i = \left(\frac{1}{n_i} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{w}_i\|^r \right)^{1/r} \tag{16.25}$$

where n_i is the number of vectors in C_i. (Compare this definition with that of the diameter of a cluster defined earlier.)

In [Pal 97] three variants of the DB index, based again on the MST, RNG, and GG concepts, are proposed. We focus on the MST case. Let s_i^{MST} be the $\text{diam}_i^{\text{MST}}$, as defined in the Dunn-like index, and let d_{ij} be the distance between the mean vectors of C_i, C_j. Then, we define

$$R_{ij}^{\text{MST}} = \frac{s_i^{\text{MST}} + s_j^{\text{MST}}}{d_{ij}} \tag{16.26}$$

It is easy to show that R_{ij}^{MST} satisfies the conditions (C1)–(C5) (see Problem 16.10). Defining $R_i^{\text{MST}} = \max_{j=1,\ldots,m,j \neq i} R_{ij}^{\text{MST}}$, the MST DB index is defined as

$$DB_m^{\text{MST}} = \frac{1}{m} \sum_{i=1}^{m} R_i^{\text{MST}} \tag{16.27}$$

The minimum in the plot of DB_m^{MST} versus m is an indication of the number of clusters that underlie X.

Using arguments similar to these, we may define DB_m^{RNG} and DB_m^{GG}.

- *The silhouette index* ([Kauf 90]). Let C_{c_i} denote the cluster where $\boldsymbol{x}_i \in X$ belongs, $i = 1, \ldots, N$. For each $\boldsymbol{x}_i$ let a_i be the average distance between $\boldsymbol{x}_i$ and the rest of the elements of C_{c_i}, that is,

$$a_i = d_{\text{avg}}^{\text{ps}}(\boldsymbol{x}_i, C_{c_i} - \{\boldsymbol{x}_i\})$$

where $d_{\text{avg}}^{\text{ps}}(\cdot,\cdot)$ denotes the average distance measure between a point and a set (see Section 11.2.1). Let also b_i be the average distance between $\boldsymbol{x}_i$ and its closest cluster excluding C_{c_i}, that is,

$$b_i = \min_{k=1,\ldots,m, k \neq c_i} d_{\text{avg}}^{\text{ps}}(\boldsymbol{x}_i, C_k)$$

Then the *silhouette width* of $\boldsymbol{x}_i$ is defined as

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{16.28}$$

It is not difficult to see that $-1 \leq s_i \leq 1$. Values of s_i close to 1 imply that the distance of $\boldsymbol{x}_i$ from the cluster where it belongs (C_{c_i}) is significantly less than the distance between $\boldsymbol{x}_i$ and its nearest cluster excluding C_{c_i}. This is an indication that $\boldsymbol{x}_i$ is well clustered. On the other hand, values of s_i close to -1 imply that the distance between $\boldsymbol{x}_i$ and C_{c_i} is significantly higher than the distance between $\boldsymbol{x}_i$ and its nearest cluster excluding C_{c_i}. This is an indication that $\boldsymbol{x}_i$ is not well clustered. Finally, values of s_i close to 0 indicate that $\boldsymbol{x}_i$ lies close to the border between the two clusters.

Based on the definition of s_i, the *silhouette of the cluster* $C_j, j = 1, \ldots, m$, is defined as

$$S_j = \frac{1}{n_j} \sum_{i:\boldsymbol{x}_i \in C_j} s_i \tag{16.29}$$

where n_j is the cardinality of C_j, and the *global silhouette index* is defined as

$$\mathcal{S}_m = \frac{1}{m} \sum_{j=1}^{m} S_j \tag{16.30}$$

Clearly, $\mathcal{S}_m \in [-1,\ 1]$. In addition, the higher the value of $\mathcal{S}_m$, the better the corresponding clustering is. Therefore, the maximum in the plot of $\mathcal{S}_m$ versus m is taken to indicate the underlying number of clusters in X.

- *The Gap statistic* ([Tibs 01]). Let D_q denote the sum of the distances between all pairs of patterns in cluster C_q, that is,

$$D_q = \sum_{\boldsymbol{x}_i \in C_q} \sum_{\boldsymbol{x}_j \in C_q} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

and let

$$W_m = \sum_{q=1}^{m} \frac{1}{2n_q} D_q \tag{16.31}$$

Clearly, a low value of W_m indicates a clustering of compact clusters.

The idea here is to compare the curve of $\log W_m$ versus m with the corresponding curve obtained from data uniformly distributed within a hyperrectangle that contains the data points of X [Hast 01] (see also [Tibs 01] for a more formal discussion on this issue). To this end, and for each m, n data sets $X_m^r, r = 1, \ldots, n$, are generated, as indicated before, and the average (in theory the expectation) $E_n(\log(W_m^r))$ over the $\log(W_m^r)$s of the corresponding X_m^rs is computed. Then the value of m for which $\log(W_m)$ falls the farthest below the reference curve formed by $E_n(\log(W_m^r))$ is taken to indicate the number of clusters in X. This is formalized via the so-called *Gap statistic*, which is defined as

$$\text{Gap}_n(m) = E_n(\log(W_m^r)) - \log(W_m) \tag{16.32}$$

The estimate of the number of clusters in X is taken to be the value that maximizes $\text{Gap}_n(m)$ (within some tolerance).

For the computational implementation of the Gap statistic we proceed as follows:

- For each value of m in $[m_{\min},\ m_{\max}]$ do
  - ○ Cluster the data set X and compute $\log(W_m)$
  - ○ Generate n reference data sets and compute the Gap statistic via Eq. (16.32).
  - ○ Define $s_m = sd_m\sqrt{1 + 1/n}$, where sd_m is the standard deviation of the $\log(W_m^r)$s around their average value.
- Choose the smallest m for which $\text{Gap}_n(m) \geq \text{Gap}_n(m+1) - s_{m+1}$

The Gap statistic can be used with any distance measure between points. In addition, it works well for the case where the data of X form a single cluster. Experimental results ([Tibs 01]) show that the Gap statistic outperforms several other indices. However, when the data are concentrated on a subspace of $\mathcal{R}^l$, the method generating the X_m^rs, as described before, degrades the performance of the Gap statistic.

- *Information Theory based criteria.* A different philosophy that may be used for the estimation of the number of clusters m relies on the determination of a model that best fits the available data, without having any knowledge of their true distribution (see, for example, [Lu 00]).

  Let us define the following criterion function:

$$C(\boldsymbol{\theta}, K) = -2L(\boldsymbol{\theta}) + \phi(K) \tag{16.33}$$

where $\boldsymbol{\theta}$ is the parameter vector of the model, $L(\boldsymbol{\theta})$, is the log-likelihood function (see Eq. (2.58)), K is the order of the model, that is, the dimensionality of $\boldsymbol{\theta}$, and ϕ is an increasing function of K. Typical choices of ϕ are $\phi(K) = 2K$ (Akaike Information Criterion, AIC [Akai 85]), $\phi(K) = \frac{2KN}{N-K-1}$ (Consistent AIC [Hurv 89]), $\phi(K) = K \ln N$ (Minimum Description Length (MDL) Criterion [Riss 78, Riss 89] and Bayesian Information Criterion (BIC) [Schw 76, Fral 98]). Note that K is a strictly increasing function of the number of clusters, m, since the higher the m, the larger the dimensionality of $\boldsymbol{\theta}$. For example, in the case where $p(\boldsymbol{x}; \boldsymbol{\theta})$ is a weighted summation of m l-dimensional Gaussian distributions, each one corresponding to a cluster, $\boldsymbol{\theta}$ consists of the ml parameters associated with the mean values of the distributions, plus $m\frac{l(l+1)}{2}$ parameters associated with the covariance matrices of the distributions plus the $m-1$ weighting parameters. Thus, $K = (l + \frac{l(l+1)}{2} + 1)m - 1$. In words, K is an increasing linear function of m.

The aim is to minimize C with respect to $\boldsymbol{\theta}$ and K. We proceed as follows. First, the set of candidate models is fixed, involving models of similar structure but of different orders. Let $m \in [m_{\min}, m_{\max}]$ for the models of the above set. Then for each value of $m_i \in [m_{\min}, m_{\max}]$, we optimize $C(\boldsymbol{\theta}, m_i)$ with respect to $\boldsymbol{\theta}$, that is, we determine the maximum likelihood estimation $\boldsymbol{\theta_i}$. Then, among all pairs $(\boldsymbol{\theta}_i, m_i)$, we choose the one, say $(\boldsymbol{\theta}_j, m_j)$, that minimizes C. Thus, the estimated number of clusters is m_j. In the case where it is desirable to choose the best among models of different structure, we first identify all the subsets, each one containing similar models of differing order. Then, we determine the best model of each subset as described above. Finally, among these models, we select the one that leads to the minimum value of C.

Other indices suitable for hard clusterings have also been proposed. For example, in [Mill 80] and [Mill 85] many indices of this kind are tested on specific data sets (see, also, [Gord 99]). Also, in [Kirl 00] two new indices are presented and their relation to the method discussed in Section 12.3, for estimating the number of clusters, is investigated. Additional indices may be found in [Shar 96, Halk 00]. In [Halk 01] an index that takes into account the density of the clusters is proposed. An evaluation of several indices may be found in [Halk 02a, Halk 02b]. Finally, in [Bout 04] a number of validity indices suitable for graph partitioning is considered.

16.4.2 Fuzzy Clustering

In this section we consider indices suitable for fuzzy clustering. In this context, we seek clusterings that are not very fuzzy, that is, those whose clusters exhibit small overlap. In other words, we seek clusterings where most of the vectors of X exhibit high grade of membership in only one cluster. Recall that a fuzzy clustering is defined by the $N \times m$ matrix $U = [u_{ij}]$, where u_{ij} denotes the grade of membership of the vector $\boldsymbol{x}_i$ in the j-th cluster. Also, let $W = \{\boldsymbol{w}_j, j = 1, \ldots, m\}$ be the set of the cluster representatives.

The strategy followed for the hard clustering case is also adopted here. That is, we define an appropriate index q (not to be confused with the fuzzifier) and we search for the minimum or the maximum in the plot of q versus m. In the case where q exhibits a trend with respect to m in the range $[m_{\min}, m_{\max}]$, we seek a significant knee of decrease or increase of q.

Indices for Clusters with Point Representatives

A. Indices that Involve Only U

One such index is the *partition coefficient* [Bezd 74], which is defined as

$$PC = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij}^2 \tag{16.34}$$

where u_{ij}'s are the values obtained after the convergence of the adopted fuzzy clustering algorithm.

The range of values for PC is $[1/m, 1]$. This index is computed for values of m greater than 1, since for $m = 1$, it is trivially equal to 1. The closer to unity the PC, the harder the clustering is or, alternatively, the smaller the "sharing" of the vectors in X among different clusters. The lowest value of PC is obtained when all u_{ij}'s are equal, that is, $u_{ij} = 1/m, j = 1, \ldots, m, i = 1, \ldots, N$. Thus, the closer the value of PC to $1/m$, the fuzzier the clustering. A value close to $1/m$ indicates that either X possesses no clustering structure or the adopted clustering algorithm failed to unravel it [Pal 95].

Another index of this category is the *partition entropy coefficient* [Bezd 75], which is defined as

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}(u_{ij}\log_a u_{ij}) \tag{16.35}$$

where a is the base of the logarithm. This index is also computed for values of m greater than 1. Its minimum value equals 0 and its maximum $\log_a m$. The closer the value of PE to 0, the harder the clustering is. On the other hand, the closer the value of PE to $\log_a m$, the fuzzier the clustering is. As in the previous case, values close to $\log_a m$ indicate the absence of any clustering structure in X or the inability of the clustering algorithm to reveal it [Pal 95].

Both of these indices measure the amount of "overlap" among clusters, without utilizing any additional information concerning the positions of the data vectors and the cluster representatives in space.

A disadvantage of both PC and PE indices is that they exhibit a dependence on m with a trend to increase or decrease, respectively, as m increases. Thus, one seeks significant knees of increase (for PC) or decrease (for PE) in the plot of the indices versus m. Moreover, they are also sensitive to the fuzzifier q. It can be shown (Problem 16.13) that as $q \to 1^+$,[6] both PC and PE give the same values for all m's;

[6] This notation means that q tends to 1 from the right.

that is, they are unable to discriminate between different values of m. On the other hand, as $q \to \infty$, both PC and PE exhibit the most significant knee at $m = 2$ (see Problem 16.13). The behavior of PC and PE is illustrated via the following example.

Example 16.8

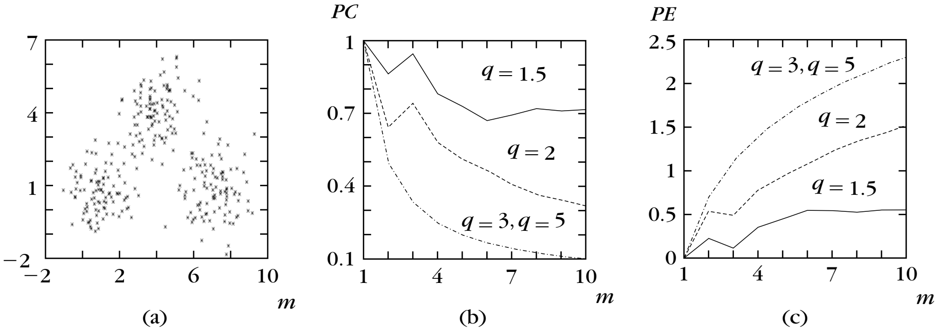
Let X be a data set that consists of three groups of two-dimensional vectors, each containing 100 vectors. The groups stem from normal distributions with means $[1, 1]^T$, $[4, 4]^T$, $[7, 1]^T$, respectively (see Figure 16.7a). All covariance matrices are equal to the identity 2×2 matrix. We run the fuzzy c-means algorithm for $q = 1.5, 2, 3, 5$ and $m = 1, \ldots, 10$. Figure 16.7b shows the behavior of the PC index. One can observe that for $q = 1.5$ and $q = 2$, the corresponding plots exhibit a significant knee at $m = 3$, which is the correct number of (the natural) clusters. The plots for $q = 3$ and $q = 5$ coincide. This implies that no significant change in the behavior of the index is expected for $q \geq 3$. Moreover, no peak is encountered; that is, no conjecture can be made for the number of clusters. Also, notice the general decreasing trend as m increases.

Figure 16.7c shows the behavior of the PE index. The plots corresponding to $q = 1.5$ and $q = 2$ exhibit a significant knee at $m = 3$. Also, as in the previous case, no significant change in the behavior of the index is expected for $q \geq 3$, and no minimum is encountered for $q \geq 3$. Finally, PE exhibits an increasing trend as m increases.

Other indices of this kind have also been proposed in the literature (e.g., [Wind 81]). We consider next indices that involve X, U, and W.

B. Indices Involving *W, U, and the Data Set X*

Let us define the so-called *cluster variation* as $\sigma_j^q = \sum_{i=1}^N u_{ij}^q \|x_i - w_j\|^2$ (compare this with the dispersion used in the DB index) and the *total variation* as $\sigma_q = \sum_{j=1}^m \sigma_j^q$. The parameter σ_q may be viewed as a measure of compactness of the specific clustering. Also let $d_{\min} = \min_{i,j=1,\ldots,m, i \neq j} \|w_i - w_j\|^2$ be a measure of



(a)

(b)

(c)

FIGURE 16.7

(a) The data set. (b) The plot of PC versus m. (c) The plot of PE versus m.

separability of the clusters in X, where $\boldsymbol{w}_j$ denotes the representative of the j-th cluster $j = 1, \ldots, m$. Then the *Xie-Beni index*, which is also called the *compactness and separation validity function*, is defined as

$$XB = \frac{\sigma_2/N}{d_{\min}} \tag{16.36}$$

This index is usually employed for the validation of clusterings produced by the fuzzy c-means algorithm when the Euclidean distance is in use. Note that despite the fact that the fuzzifier q in the fuzzy c-means may have any value greater than 1, *in the XB index the value of q involved in σ_q is restricted to* 2.

It is clear that for compact and well-separated clusters, small values of XB are expected. On the other hand, small values of XB indicate compact and well-separated clusters. As stated in [Xie 91], the XB index decreases monotonically as m gets very close to N. One way to handle this problem is to determine the starting point, $m_{\max}$, of the monotonicity behavior and to search for the minimum value of XB in the range $[2, m_{\max}]$.

Let

$$J_q = \sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij}^q \|\boldsymbol{x}_i - \boldsymbol{w}_j\|^2 \tag{16.37}$$

(recall that this is the cost function minimized by the fuzzy c-means clustering algorithm when the squared Euclidean distance is in use). Then XB may be written in terms of J_2 as

$$XB = \frac{J_2}{N d_{\min}} \tag{16.38}$$

Thus, minimization of XB implies minimization of J_2.

Removal of the constraint $q = 2$, used in the definition of XB, allows the definition of the *generalized XB index* as

$$XB_q = \frac{\sigma_q}{N d_{\min}} \tag{16.39}$$

It can be shown (Problem 16.14) that, as $q \to \infty$, XB tends to ∞ and XB_q becomes indeterminate.

Another index that combines X, W, and U is the *Fukuyama-Sugeno index* [Pal 95], which is defined as

$$FS_q = \sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij}^q \left(\|\boldsymbol{x}_i - \boldsymbol{w}_j\|_A^2 - \|\boldsymbol{w}_j - \boldsymbol{w}\|_A^2\right) \tag{16.40}$$

where $\boldsymbol{w}$ is the mean vector of X and A is an $l \times l$ positive definite, symmetric matrix. Recall that $\|\cdot\|_A$ is the A norm defined in Section 2.4. When $A = I$, the above distance becomes the squared Euclidean distance.

The first of the two terms in the parenthesis measures the compactness of the clusters, and the second one measures the distance of the cluster representatives,

$\boldsymbol{w}_i$, from the overall mean vector $\boldsymbol{w}$. It is clear that for compact and well-separated clusters we expect small values for FS_q. Furthermore, small values of FS_q are indicative of compact and well-separated clusters. As far as the limiting behavior of FS_q is concerned, it can be shown (Problem 16.15) that (a) as $q \to 1^+$, FS_q behaves like $tr(S_w)$, the trace of the within scatter matrix (see Chapter 5), and (b) as $q \to \infty$, FS_q tends to 0.

In [Gath 89], three additional indices are proposed that are based on the concepts of hypervolume and density. Let us define the *fuzzy covariance matrix* of the j-th cluster as

$$\Sigma_j = \frac{\sum_{i=1}^{N} u_{ij}^q (\boldsymbol{x}_i - \boldsymbol{w}_j)(\boldsymbol{x}_i - \boldsymbol{w}_j)^T}{\sum_{i=1}^{N} u_{ij}^q} \tag{16.41}$$

The *fuzzy hypervolume* of the j-th cluster is defined as

$$V_j = |\Sigma_j|^{1/2} \tag{16.42}$$

where $|\Sigma_j|$ is the determinant of Σ_j. Note that this is a measure of compactness of the j-th cluster. The smaller the value of V_j, the more "compact" the j-th cluster is.

The *total fuzzy hypervolume* is defined as

$$FH = \sum_{j=1}^{m} V_j \tag{16.43}$$

Small values of FH indicate the existence of compact clusters.

Let $X_j = \{\boldsymbol{x} \in X : (\boldsymbol{x} - \boldsymbol{w}_j)^T \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{w}_j) < 1\}$; that is, X_j contains all the vectors in X that are within a prespecified (small) region around $\boldsymbol{w}_j$. Also let $S_j = \sum_{\boldsymbol{x}_i \in X_j} u_{ij}$ be the so-called sum of central members of the jth cluster. The quantity S_j/V_j is called the *fuzzy density* of the jth cluster. Then the *average partition density* is defined as

$$PA = \frac{1}{m} \sum_{j=1}^{m} \frac{S_j}{V_j} \tag{16.44}$$

A different measure is the *partition density* index and it is defined as

$$PD = \frac{S}{FH} \tag{16.45}$$

where $S = \sum_{j=1}^{m} S_j$.

"Compact" clusters lead to large values of PA and PD, and vice versa, large values of PA and PD are indications of "compact" clusters.

Note that all these indices, except PE and PC, may be used in the framework of hard clustering as well, by defining

$$u_{ij} = \begin{cases} 1, & \text{if } d(\boldsymbol{x}_i, C_j) = \min_{k=1,\ldots,m} d(\boldsymbol{x}_i, C_k) \\ 0, & \text{otherwise} \end{cases} \qquad i = 1, \ldots, N \tag{16.46}$$

Additional indices for fuzzy clustering validation are discussed in [Boug 04, Sent 07].

Indices for Shell-Shaped Clusters

Let us now focus on the case of shell-shaped clusters (see Chapter 14). The *PE* and *PC* indices, discussed previously, may also be used in this case, since they involve no information concerning the geometrical characteristics of X.[7] However, the rest of the previously discussed indices need to be modified accordingly. Here, the representatives of each cluster, are shell shaped and they are denoted by β_j. The parameter vector $\boldsymbol{\theta}_j$ contains all the necessary parameters for the identification of β_j. For a vector $\boldsymbol{x}_i$, we define its distance from β_j in terms of

$$\tau_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j^i \tag{16.47}$$

where $\boldsymbol{x}_j^i$ is the point on β_j that is, closer to $\boldsymbol{x}_i$. It is not difficult to show (Problem 16.16) that for clusters of spherical shape, where $\boldsymbol{\theta}_j = (\boldsymbol{c}_j, r_j)$, with $\boldsymbol{c}_j$ being the center and r_j being the radius of the corresponding sphere,

$$\tau_{ij} = (\boldsymbol{x}_i - \boldsymbol{c}_j) - r_j \frac{\boldsymbol{x}_i - \boldsymbol{c}_j}{\|\boldsymbol{x}_i - \boldsymbol{c}_j\|} \tag{16.48}$$

However, for general types of shells, computation of the τ_{ij}'s is not always an easy task. In such cases we resort to approximations of $\boldsymbol{x}_j^i$ ([Kris 95a]).

The *fuzzy shell covariance matrix* for the j-th cluster is defined in accordance with Eq. (16.41) as

$$\Sigma_j^S = \frac{\sum_{i=1}^N u_{ij}^q \tau_{ij} {\tau_{ij}}^T}{\sum_{i=1}^N u_{ij}^q} \tag{16.49}$$

Then the *shell hypervolume* of a cluster is defined as

$$V_j^S = |\Sigma_j^s|^{1/2} \tag{16.50}$$

Let us define $X_j^S = \{\boldsymbol{x}_i \colon {\tau_{ij}}^T (\Sigma_j^S)^{-1} \tau_{ij} < 1\}$ and $S_j^S = \sum_{\boldsymbol{x}_i \in X_j^S} u_{ij}$. Then, the *fuzzy shell density*, the *average partition shell density*, and the *shell partition density* are defined as before.

Finally, another measure suitable for shell-shaped clusters is the *total fuzzy average shell thickness*, T^S [Kris 95b], which is defined as

$$T^S = \sum_{j=1}^m T_j^S \tag{16.51}$$

where T_j^S is the so-called *fuzzy average shell thickness* of the *j*th cluster, defined as

$$T_j^S = \frac{\sum_{i=1}^N u_{ij}^q \|\tau_{ij}\|^2}{\sum_{i=1}^N u_{ij}^q} \tag{16.52}$$

[7] Such characteristics may concern the shape of the clusters, the position of the representatives, etc.

It is clear that the "thicker" the clusters, the smaller the value of T^S. Furthermore, small values of T^S indicate "thick" clusters. However, T^S tends to decrease monotonically as the number of clusters increases.

The comments made for the total fuzzy hypervolume, the average partition density and the partition density indices are also valid here.

Note that PA^S, PD^S, *and* T^S *can be thought as measures of the density of the clusters formed by the vectors of X around their representatives.*

A few other indices of this kind have also been proposed and discussed in [Dave 90, Kris 93]. A detailed overview of objective structural validity criteria is given in [Halk 02a, Halk 02b]. A general comment applied to all these indices is that they are sensitive to the size and the density of the points in the clusters.

Finally, using Eq. (16.46), we obtain the shell density, the average partition shell density, and the shell partition density for the hard clustering case.

Remarks

- An alternative way of determining the number of clusters underlying in the data set X, is to perform the so called *progressive clustering method* (e.g., [Kris 95b]). According to this method we run first the clustering algorithm at hand for an overspecified number of clusters, m. Then, we remove spurious clusters, we merge compatible clusters and we identify the "good" clusters. Let k be the number of spurious and "good" clusters defined above. Then, we temporarily remove the vectors contained in the above clusters from the data set and we apply the algorithm on the reduced data set for $m-k$ clusters. This procedure is repeated until no "good" clusters can be removed anymore or no vectors are left in the data set. The output of the above method is the set of the "good" clusters determined above.

  The advantage of this method is that, in general, it is not necessary to run the clustering algorithm for all values of m in a prespecified range. Also, this method is less influenced by the presence of noise. However, one must establish criteria concerning the merging and the removing operations involved in the above method as well as criteria for the identification of "good" clusters.

- A different philosophy for the determination of the number of clusters underlying in the data set X employs the idea of information criteria (IC), such as Akaike and the Minimum Description Length (MDL) criteria (see, e.g., [Sclo 87, Lang 98]).

16.5 VALIDITY OF INDIVIDUAL CLUSTERS

There are two cases in which individual cluster validity may be of interest. One is when we want to test whether a given subset of X forms a "good" cluster. "Good" in this case is interpreted in terms of compactness, with respect to its own data, and isolation with respect to the other vectors of X. The other case concerns the validation of a cluster resulting from the application of a clustering algorithm. To this end, both external and internal criteria may be used.

16.5.1 External Criteria

In this section we consider hard clusters and ordinal-type proximity matrices [Bail 82, Jain 88]. The goal is to test whether a given subset of X forms a compact and well-separated cluster. In [Bail 82], two indices are defined, one for compactness and one for isolation. Both are based on graph theory concepts. However, some necessary definitions must first be provided.

Let us consider the proximity graph $G(p)$, with p ($<N(N-1)/2$) edges, whose vertices correspond to the N vectors of X and whose edges correspond to the p smallest entries of the upper diagonal of the proximity matrix of X, P. In other words, a pair of vertices $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is connected with an edge if the dissimilarity $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is among the p smallest dissimilarity values of all possible pairs of vectors in X (see Chapter 13). Also, let C be a predetermined subset of X, with k vectors. Our goal is to determine whether C is a good cluster. For the $G(p)$ and the given P, we define the sets $A_{\text{in}}(p)$, $A_{\text{out}}(p)$, and $A_{\text{bet}}(p)$ as follows: (a) $A_{\text{in}}(p)$ is the set of edges whose end points are vectors in C, (b) $A_{\text{out}}(p)$ is the set of edges whose end points are vectors in $X-C$, and (c) $A_{\text{bet}}(p)$ is the set of edges that connect vectors in C with vectors in $X-C$.

For a given $G(p)$, let $q_C(p)$ be the number of edges connecting vertices in C with vertices in $X-C$ and $r_C(p)$ be the set of edges connecting vertices in C. Clearly, these indices depend on p. It is easy to see that low values of $q_C(p)$ indicate a well-isolated cluster, and large values of $r_C(p)$ indicate a compact cluster. In order to extract conclusions about the compactness and isolation of C, we consider the behavior of these indices with respect to p. To this end, we plot the indices versus p. It is expected that an isolated and compact cluster will exhibit low values for $q_C(p)$ and high values for $r_C(p)$, for a "wide" range of values of p. The size of this range is application dependent.

A drawback of these indices is that they do not provide information with respect to a random population. To overcome this, an extension of the indices within the probabilistic framework is discussed in [Bail 82].

16.5.2 Internal Criteria

The aim here is to validate a single cluster that results from a clustering algorithm using only the information residing in the proximity matrix, P, of X.

- *Hard clustering case*
  - *Ordinal proximity matrices.* A method for the evaluation of a cluster is given in [Ling 72] and [Ling 73]. This method is well suited for hierarchies of clusterings, produced by a hierarchical clustering algorithm. It relies on

the lifetime, $L(C)$, of a cluster C, which is given by $L(C) = d_a(C) - d_f(C)$ where $d_f(C)$ is the level of hierarchy where C is formed and $d_a(C)$ is the level where C is absorbed in a larger cluster. The statistical index used is the so-called Ling index, which is defined as the probability of the lifetime of a randomly selected cluster exceeding $L(C)$. Finally, other methods in this category are the so-called *best case method* [Bail 82] and the *CM (clustering method) reachable method* [Bake 76].

- *Ratio-scaled proximity matrices.* In this case, we may adopt the hard hypervolume and the hard density (Section 16.4.2) when we seek compact clusters and the hard shell hypervolume and hard shell density when shell-shaped clusters are considered. Here, an empirically established threshold, ε, is used and, according to whether the value of the index is greater or less than ε, C is characterized as a "good" cluster or not.

- *Fuzzy clustering case.* We first focus on shell-shaped clusters. In this context, "good" clusters are those that are *"compact" around their shell representatives*. In this framework, one can use the shell hypervolume, the shell density indices, and the fuzzy average shell thickness, defined in Section 16.4. Based on these indices, a cluster is characterized as a good one according to whether the value of the corresponding index is greater or less than a prespecified threshold ε.

All these indices do not take into account the fact that shell clusters lie in subspaces of the vector space [Kris 95b]. A criterion that takes this observation into account is the *surface density* criterion. We present the two-dimensional case. The criterion measures the number of points in a cluster per unit curve length. Let us define X' as the set of the vectors in C that lie at a distance smaller than or equal to $\tau_{\max}$ from the shell representative β of C and let $S = \sum_{j:\boldsymbol{x}_j \in X'} u_j$. Then, the surface density δ of a cluster C is defined as

$$\delta = \frac{S}{2\pi r_{\text{eff}}} \tag{16.53}$$

where r_{eff} is defined as

$$r_{\text{eff}} = \sqrt{\text{trace}\{\Sigma\}} \tag{16.54}$$

where Σ is given in Eq. (16.49) and $tr(\Sigma)$ is the trace of Σ. The quantity $2\pi r_{\text{eff}}$ may be viewed as an estimate of the arc length of C (see Problem 16.17). The higher the value of δ, the more dense the cluster is expected to be. Consider for example Figure 16.8. The clusters depicted there have a circular shape and their representative circles are of equal radius. Also, the one on the right is denser around its representative than the one on the left. The value of δ for the right cluster is greater than that for the left cluster.

For compact clusters, indices such as the fuzzy hypervolume or the fuzzy density of a cluster can be employed.

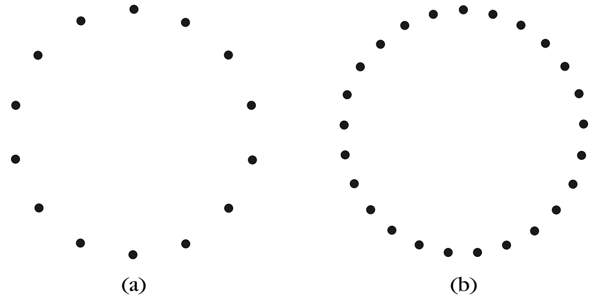


FIGURE 16.8

A sparse and a dense circular cluster.

16.6 CLUSTERING TENDENCY

As discussed in the introduction of the chapter, almost all the clustering algorithms introduced in the previous sections share an annoying feature. That is, they impose a clustering structure on a data set X even though the vectors of X do not exhibit such a structure. Thus, in order to prevent a misleading interpretation of the structure of the data set X, it would be more sensible to check first whether X possesses a clustering structure. If this is the case, then one may proceed by applying a clustering algorithm to X. Otherwise, cluster analysis is likely to lead to misleading results. The problem of determining the presence or the absence of a clustering structure in X is called *clustering tendency*. Usually, this task relies on statistical tests.

Clustering tendency methods have been applied in various application areas (e.g., [Digg 83, Ripl 81]). However, most of these methods are suitable only for $l = 2$. In the sequel, we discuss the problem in the general $l \geq 2$ case. Furthermore, we focus on methods that are suitable for detecting compact clusters (if any).

In this framework, we test the randomness (null) hypothesis (H_0) against the clustering hypothesis and the regularity hypothesis. Let us define these terms more precisely.

- "The vectors of X are randomly distributed, according to the uniform distribution in the sampling window[8] of X" (H_0).
- "The vectors of X are regularly spaced in the sampling window." This implies that, they are not too close to each other.
- "The vectors of X form clusters."

If the randomness or the regularity hypothesis is accepted, methods alternative to clustering analysis should be used for the interpretation of the data set X.

[8] In [Smit 84] the sampling window is mathematically defined as the compact convex support set for the underlying distribution of the vectors of the data set X.

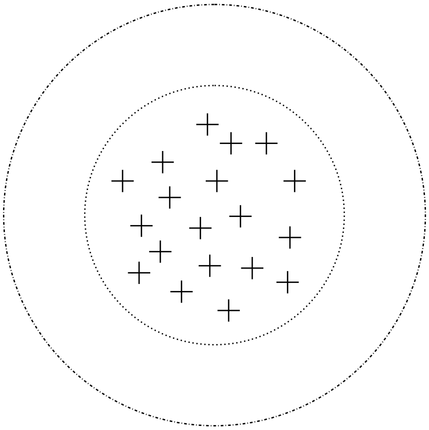


FIGURE 16.9

See text for explanation.

There are two key points that have an important influence on the performance of many statistical tests used in clustering tendency. The first is the dimensionality of the data, l, which affects the performance in a nonobvious way. This dependence can be revealed through simulations [Pana 83].

The other key point is the sampling window. Apart from artificial experiments, in practice, we do not know the sampling window. One of the problems that this may cause is demonstrated in Figure 16.9. The vectors in the dashed circle are uniformly distributed in it. Thus, we expect that tests for randomness will identify this situation. However, if we use as sampling window the region surrounded by the dash-dotted line (for the same data set), the vectors are no longer uniformly distributed and the tests for randomness may fail to accept H_0. Moreover, due to the finite extent of the window, the statistical characteristics of the data are different near the edges of the sampling window than they are in its center. For example, the distribution of the distances of a vector $\boldsymbol{x} \in X$ from the rest of the vectors of X is different when $\boldsymbol{x}$ is in the center than when it is near the border of the sampling window. One way to overcome this situation is to use a periodic extension of the sampling window. Another popular technique is to consider data in a smaller area inside the sampling window, known as *sampling frame*. With this method, we overcome the boundary effects in the sampling frame by considering points outside it and inside the sampling frame, for the estimation of statistical properties.

Example 16.9

Consider a data set X that consists of 100 vectors uniformly distributed in the H_2 hypercube (see Figure 16.10a). Figure 16.10b shows the distribution of the distances between the point $\boldsymbol{x} = [0.5045, 0.4764]^T$ and each of the points of $X - \{\boldsymbol{x}\}$. Also, Figure 16.10c shows the

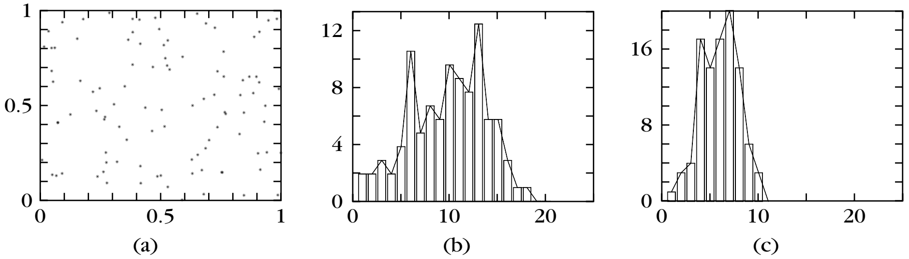


FIGURE 16.10

(a) The data set X. (b) The distribution of the distances of the point $[0.5045, 0.4764]^T$ from the remaining points in X. (c) The distribution of the distances of the point $[0.0159, 0.8089]^T$ from the remaining points in X.

distribution of the distances between the point $\boldsymbol{y} = [0.0159,\ 0.8089]^T$ and each of the points of $X - \{\boldsymbol{y}\}$. Note that $\boldsymbol{x}$ lies close to the center of H_2 and $\boldsymbol{y}$ lies close to its border.

A method for estimating the sampling window is to use the convex hull of the vectors in X. However, the distributions for the tests, derived using this sampling window, depend on the specific data at hand. A second drawback associated with this approach is the high computational cost for computing the convex hull of X. An alternative [Zeng 85, Dube 87b] that seems to work well in practice is to define the sampling window as the hypersphere centered at the mean point of X and including half of its vectors. The fact that half of the vectors are discarded is not so crucial, because in the current framework we want to test only whether the vectors of X possess a clustering structure. If this is the case, then the clusters will be identified by applying a clustering algorithm to all the data of X. Notice the similarity to the sampling frame technique discussed earlier.

In the sequel, we define various test statistics, q, suitable for the detection of clustering tendency. Recall that a crucial quantity we have to determine is $p(q|H_0)$. Moreover, in order to measure the power of q against the regularity and the clustering tendency hypotheses, we also need to determine the respective pdf's under these hypotheses. In the sequel, we provide general guidelines on how to generate clustered and regularly spaced data sets. This is required in order to estimate the pdf's of q under regularity and clustering tendency hypotheses, via Monte Carlo simulations. Randomly spaced data sets may be generated by inserting vectors in the sampling window, according to the uniform distribution.

- *Generation of clustered data.* A well-known procedure for generating (compact) clustered data is the Neyman–Scott procedure [Neym 72]. This procedure assumes that the sampling window is known. It produces a random number of compact clusters, formed at random positions in the sampling window and each consisting of a random number of points. The number of points in each cluster follows the Poisson distribution (Appendix A). The technique requires as inputs the total number of points N of the set, the intensity of the Poisson process λ, and the spread parameter σ that controls the spread of each cluster around its center. According to this procedure, we randomly insert a point $\boldsymbol{y}_i$ in the sampling window, following the uniform distribution. This point serves as the center of the ith cluster, and we determine its number of vectors, n_i, using the Poisson distribution. Then the n_i points around $\boldsymbol{y}_i$ are generated according to the normal distribution with mean $\boldsymbol{y}_i$ and covariance matrix $\sigma^2 I$. If a point turns out to be outside the sampling window, we ignore it and another one is generated. This procedure is repeated until N points have been inserted in the sampling window (see Figures 16.11a and b). In some cases, $\boldsymbol{y}_i$'s are also included as vectors in the set.

- *Generation of regularly spaced data.* Perhaps the simplest way to produce regularly spaced points is to define a lattice in the convex hull of X and to place the vectors at its vertices. An alternative procedure, known as *simple sequential inhibition (SSI)* (see, e.g., [Jain 88, Zeng 85]), is the following. The points $\boldsymbol{y}_i$ are inserted in the sampling window one at a time. For each point we define a hypersphere of radius r centered at $\boldsymbol{y}_i$. The next point can be placed anywhere in the sampling window in such a way that its hypersphere does not intersect with any of the hyperspheres defined by the previously inserted points. The procedure stops when a predetermined number of points have been inserted in the sampling window, or when no more points can be inserted in the sampling window, after say a few thousand trials (see Figure 16.11c). A variation of this model allows intersection of these hyperspheres up to a certain degree. A measure of the degree of fulfillment of the sampling window is the so-called *packing density*, which is defined as

$$\rho = \frac{L}{V} V_r \qquad (16.55)$$

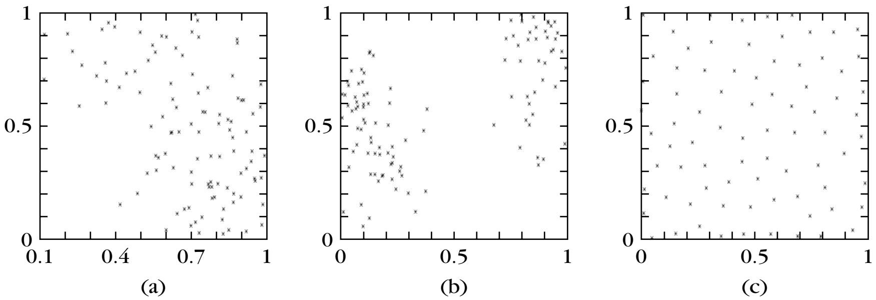


FIGURE 16.11

(a) and (b) Clustered data sets produced by the Neyman–Scott process. (c) Regularly spaced data produced by the SSI model.

where L/V is the average number of points per unit volume and V_r is the volume of a hypersphere of radius r. V_r can be written as

$$V_r = Ar^l \tag{16.56}$$

where A is the volume of the l-dimensional hypersphere with unit radius, which is given by

$$A = \frac{\pi^{l/2}}{\Gamma(l/2+1)} \tag{16.57}$$

and $\Gamma(\cdot)$ is the gamma function (Appendix A).

16.6.1 Tests for Spatial Randomness

Several tests for spatial randomness have been proposed in the literature. All of them assume knowledge of the sampling window. The *scan test* ([Naus 82, Cono 79]), the *quadrat analysis* [Grei 64, Piel 69, Mead 74], the *second moment structure* [Ripl 77], and the *interpoint distances* [Ripl 78, Silv 78, Stra 75] provide us with tests for clustering tendency that have been extensively used when $l = 2$. In the sequel, we discuss three methods for determining clustering tendency that are well suited for the general $l \geq 2$ case. All these methods require knowledge of the sampling window.

Tests Based on Structural Graphs

In this section, we discuss a test for testing randomness, that is, based on the idea of the minimum spanning tree (MST) ([Smit 84]). First, we determine the convex region where the vectors of X lie. Then, we generate M vectors that are uniformly distributed over a region that approximates the convex region found before (usually $M = N$). These vectors constitute the set X'. Next we find the MST of $X \cup X'$ and we determine the number of edges, q, that connect vectors of X with vectors of X'. This number is used as the statistic index. If X contains clusters, then we expect q to be small. Conversely, small values of q indicate the presence of clusters. On the other hand, large values of q indicate a regular arrangement of the vectors of X.

Let e be the number of pairs of the MST edges that share a node. In [Frie 79], the following expressions for the mean value of q and the variance of q under the null (randomness) hypothesis, conditioned on e, are derived:

$$E(q|H_0) = \frac{2MN}{M+N} \tag{16.58}$$

and

$$\begin{aligned}\text{var}(q|e, H_0) = \frac{2MN}{L(L-1)}\Bigg[&\frac{2MN-L}{L} \\ &+ \frac{e-L+2}{(L-2)(L-3)}[L(L-1)-4MN+2]\Bigg]\end{aligned} \tag{16.59}$$

where $L = M + N$. Moreover, it can be shown [Frie 79] that if $M, N \rightarrow \infty$ and M/N is away from 0 and ∞, the pdf of the statistic

$$q' = \frac{q - E(q|H_0)}{\sqrt{\text{var}(q|e, H_0)}} \tag{16.60}$$

is approximately given by the standard normal distribution. Thus, we reject H_0 at significance level ρ if q' is less than the ρ-percentile of the standard normal distribution. This test exhibits high power against clustering tendency and little power against regularity [Jain 88].

Tests Based on Nearest Neighbor Distances

Two tests of this kind are the Hopkins test [Hopk 54] and the Cox–Lewis test [Cox 76, Pana 83]. The tests rely on the distances between the vectors of X and a number of vectors which are randomly placed in the sampling window.

The Hopkins Test

Let $X' = \{\boldsymbol{y}_i, i = 1, \ldots, M\}, M \ll N$,[9] be a set of vectors that are randomly distributed in the sampling window, following the uniform distribution. Also let $X_1 \subset X$ be a set of M randomly chosen vectors of X. Let d_j be the distance from $\boldsymbol{y}_j \in X^{'}$ to its closest vector in X_1, denoted by $\boldsymbol{x}_j$, and δ_j be the distance from $\boldsymbol{x}_j$ to its closest vector in $X_1 - \{\boldsymbol{x}_j\}$. Then the Hopkins statistic involves the lth powers of d_j and δ_j and it is defined as [Jain 88]

$$h = \frac{\sum_{j=1}^{M} d_j^l}{\sum_{j=1}^{M} d_j^l + \sum_{j=1}^{M} \delta_j^l} \tag{16.61}$$

This statistic compares the nearest neighbor distribution of the points in X_1 with that from the points in $X^{'}$. When X contains clusters, the distances between nearest neighbor points in X_1 are expected to be small, on the average, and, thus, large values of h are expected. Furthermore, large values of h indicate the presence of a clustering structure in X. When the points in X are regularly distributed in the sampling window, it is expected that, on the average, the term $\sum_{j=1}^{M} d_j^l$ is smaller than $\sum_{j=1}^{M} \delta_j^l$, thus leading to small values of h. Also, small values of h indicate the presence of regularly spaced points. Finally, a value around 1/2 is an indication that the vectors of X are randomly distributed over the sampling window. It can be shown (e.g., [Jain 88]) that if the generated vectors are distributed according to a Poisson random process (hypothesis of randomness) and all nearest neighbor distances are statistically independent, h (under H_0) follows a beta distribution, with (M, M) parameters (Appendix A).

[9] Typically $M = 0.1N$.

Simulation results [Zeng 85] show that this test exhibits high power against regularity for a hypercubic sampling window and periodic boundaries, for $l = 2, \ldots, 5$. However, its power is limited against clustering tendency.

The Cox–Lewis Test

This test is less intuitive than the previous one. It was first proposed in [Cox 76] for the two-dimensional case and it has been extended to the general $l \geq 2$ dimensional case in [Pana 83]. It follows the setup of the previous test with the exception that X_1 need not be defined. For each $\boldsymbol{y}_j \in X'$, we determine its closest vector in X, say $\boldsymbol{x}_j$, and then we determine the vector closest to $\boldsymbol{x}_j$ in $X - \{\boldsymbol{x}_j\}$, say $\boldsymbol{x}_i$. Let d_j be the distance between $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$ and δ_j the distance between $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$. We consider all $\boldsymbol{y}_j$'s for which $2d_j/\delta_j$ is greater than or equal to one. Let $M^{'}$ be the number of such $\boldsymbol{y}_j$'s. Then, an appropriate function R_j of $2d_j/\delta_j$ (see [Pana 83]) is defined for these $\boldsymbol{y}_j$'s. Finally, we define the statistic

$$R = \frac{1}{M'} \sum_{j=1}^{M'} R_j \tag{16.62}$$

It can be shown [Pana 83] that R, under H_0, has an approximately normal distribution with mean 1/2 and variance $12M'$. Small values of R indicate the presence of a clustering structure in X, and large values indicate a regular structure in X. Finally, values around the mean of R indicate that the vectors of X are randomly arranged in the sampling window. Simulation results [Zeng 85] show that the Cox–Lewis test exhibits inferior performance compared with the Hopkins test against the clustering alternative. However, this is not the case against the regularity hypothesis.

Two additional tests are the so called T-squared sampling tests, introduced in [Besa 73]. However, simulation results [Zeng 85] show that the these two tests exhibit rather poor performance compared with the Hopkins and Cox–Lewis tests.

A Sparse Decomposition Technique

This technique begins with the data set X and sequentially removes vectors from it until no vectors are left [Hoff 87]. Before we proceed further, some definitions are needed. A *sequential decomposition D* of X is a partition of X into $L_1, \ldots, L_k$ sets, such that the order of their formation matters. L_i's are also called *decomposition layers*.

We denote by $MST(X)$ the MST corresponding to X. Let $S(X)$ be the set derived from X according to the following procedure. Initially, $S(X) = \emptyset$. We move an end point $\boldsymbol{x}$ of the longest edge, e, of the $MST(X)$ to $S(X)$. Also, we mark this point and all points that lie at a distance less than or equal to b from $\boldsymbol{x}$, where b is the length of e. Then, we determine the unmarked point, $\boldsymbol{y} \in X$, that lies closer to $S(X)$ and we move it to $S(X)$. Also, we mark all the unmarked vectors that lie at a distance no greater than b from $\boldsymbol{y}$. We apply the same procedure for all the unmarked vectors of X. The procedure terminates when all vectors are marked.

Let us define $R(X) \equiv X - S(X)$. Setting $X = R^0(X)$, we define

$$L_i = S(R^{i-1}(X)), \qquad i = 1, \ldots, k \tag{16.63}$$

where k is the smallest integer such that $R^k(X) = \emptyset$. The index i denotes the so-called *decomposition layer*. Intuitively speaking, the procedure sequentially "peels" X until all of its vectors have been removed.

The information that becomes available to us after the application of the decomposition procedure is (a) the number of decomposition layers k, (b) the decomposition layers L_i, (c) the cardinality, l_i, of the L_i decomposition layer, $i = 1, \ldots, k$, and (d) the sequence of the longest MST edges used in deriving the decomposition layers. The decomposition procedure gives different results when the vectors of X are clustered and when they are regularly spaced or randomly distributed in the sampling window. Based on this observation we may define statistical indices utilizing the information associated with this decomposition procedure. For example, it is expected that the number of decomposition layers, k, is smaller for random data than it is for clustered data. Also, it is smaller for regularly spaced data than for random data (see Problem 16.20). This situation is illustrated in the following example.

Example 16.10

(a) We consider a data set X_1 of 60 two-dimensional points in the unit square. The first 15 points stem from a normal distribution, with mean $[0.2, 0.2]^T$ and covariance matrix $0.15I$. The second, the third, and the fourth group of 15 points also stem from normal distributions with means $[0.2, 0.8]^T$, $[0.8, 0.2]^T$, and $[0.8, 0.8]^T$, respectively. Their covariance matrices are also equal to $0.15I$. Applying the sparse decomposition technique on X_1, we obtain 15 decomposition layers.

(b) We consider another data set X_2 of 60 two-dimensional points, which are now randomly distributed in the unit square. The sparse decomposition technique in this case gives 10 decomposition layers.

(c) Finally, we generate a data set X_3 of 60 two-dimensional points regularly distributed in the unit square, using the simple sequential inhibition (SSI) procedure. The sparse decomposition technique gives 7 decomposition layers in this case.

Figures 16.12, 16.13, and 16.14 show the first four decomposition layers for clustered, random, and regularly spaced data. It is clear that the rate of point removal is much slower for the clustered data and much faster for the regular data.

Several tests that rely on the preceding information are discussed in [Hoff 87]. One such statistic that exhibits good performance is the so-called *P statistic*, which is defined as follows:

$$P = \prod\nolimits_{i=1}^{k} \frac{l_i}{n_i - l_i} \tag{16.64}$$

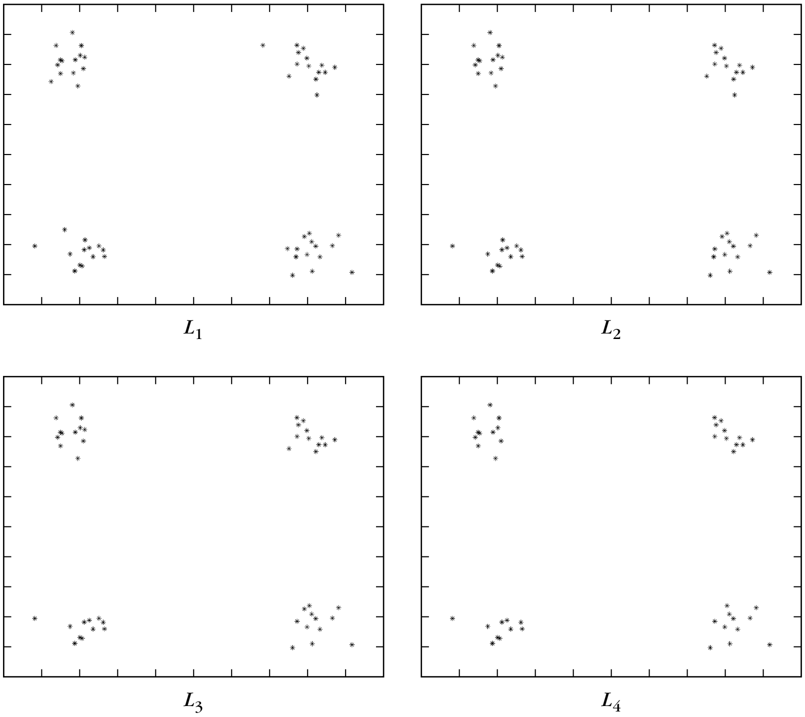


FIGURE 16.12

The first four decomposition layers for clustered data in the unit square (Example 16.10(a)).

where n_i is the number of points in $R^{i-1}(X)$. In words, each factor of P is the ratio of the removed to the remaining points at each decomposition stage.

Preliminary simulation results show high power of P against the clustering alternative. The required pdf's of P under H_0, H_1, and H_2 are estimated using Monte Carlo techniques, since it is difficult to derive theoretical results [Hoff 87].

Finally, tests for clustering tendency for the cases in which ordinal proximity matrices are in use have also been proposed (e.g., [Fill 71, Dube 79]). Most of them are based on graph theory concepts. Let $G_N(v)$ be a threshold graph with N vertices, one for each vector of X (Chapter 13). Then, graph properties, such as the node degree and the number of edges needed for $G_N(v)$ to be connected, are used in order to investigate the clustering tendency of X. Specifically, suppose that we use the number of edges n needed to make $G_N(v)$ connected. Obviously, n depends directly on v. That is, increasing v, we also increase n. Let v^* be the smallest value of v for which $G_N(v^*)$ becomes connected, for the given proximity matrix. Let V be the random variable that models v. Also, let $P(V \leq v|N)$ be the probability that a graph with N nodes and v randomly inserted edges is connected (this is provided from tables in [Ling 76]). Then, for the specific v^*, we determine $P(V \leq v^*|N)$. Very high values of $P(V \leq v^*|N)$ indicate that the proximity matrix

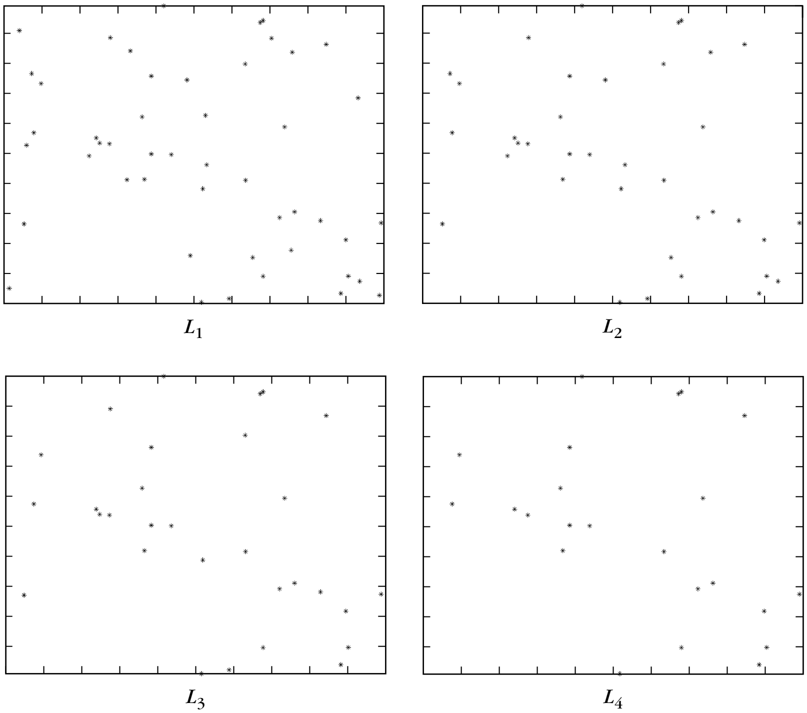


FIGURE 16.13

The first four decomposition layers for randomly distributed data in the unit square (Example 16.10(b)).

was not chosen at random. This is because the within-cluster edges will tend to occur before the between-cluster edges when the data are clustered, thus, delaying the formation of a connected graph.

16.7 PROBLEMS

16.1 Let X be a set of vectors. Show that if the number of clusters in a clustering $\mathcal{C}$ of X is m and the number of groups in a partition $\mathcal{P}$ of X is $q \neq m$, then the maximum values of the Rand, the Jaccard, and the Fowlkes and Mallows statistics are less than 1.

16.2 Prove Eq. (16.10).

16.3 **a.** Repeat Example 16.2 with two-dimensional vectors steming from the normal distributions with means $[0.2, 0.2]^T$, $[0.2, 0.8]^T$, $[0.8, 0.2]^T$, $[0.8, 0.8]^T$, and covariance matrices $0.2^2 I$.

b. Repeat the experiment when all covariance matrices are equal to $0.5^2 I$.

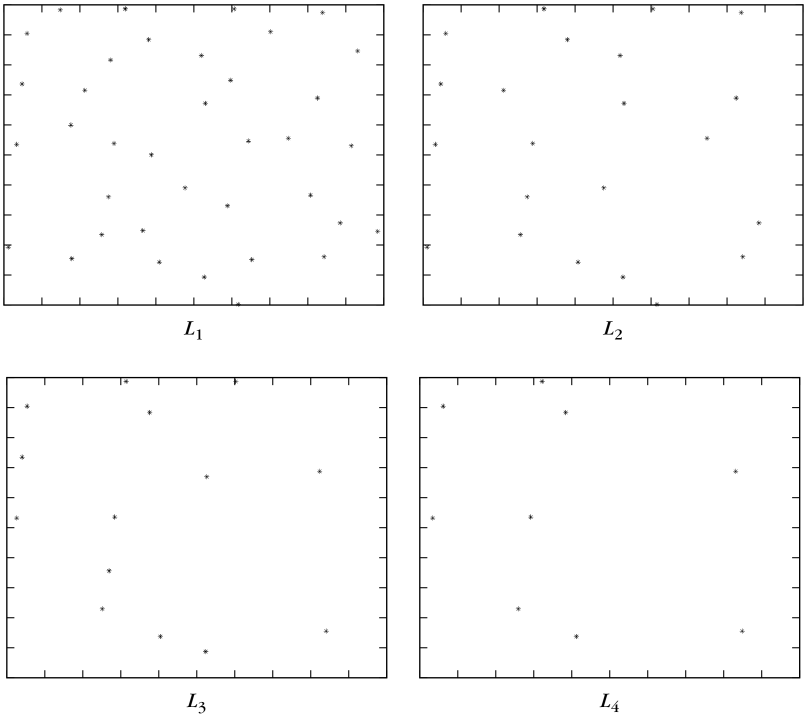


FIGURE 16.14

The first four decomposition layers for regularly spaced data in the unit square (Example 16.10(c)).

16.4 Prove that the values of the *CPCC* in Section 16.3.2 lie in the interval $[-1, 1]$.

16.5 Consider a data set X of six vectors, whose (ordinal) proximity matrix is

$$P = \begin{bmatrix} 0 & 1 & 5 & 7 & 8 & 9 \\ 1 & 0 & 3 & 6 & 10 & 11 \\ 5 & 3 & 0 & 12 & 13 & 14 \\ 7 & 6 & 12 & 0 & 6 & 4 \\ 8 & 10 & 13 & 6 & 0 & 2 \\ 9 & 11 & 14 & 4 & 2 & 0 \end{bmatrix}$$

Run the single and the complete link algorithms on X and compare the resulting dendrograms, using the γ statistic. Comment on the results.

16.6 Let $X = \{\boldsymbol{x}_i, i = 1, \dots, 12\}$, with $\boldsymbol{x}_1 = [-4, 0]^T$, $\boldsymbol{x}_2 = [-3, 1]^T$, $\boldsymbol{x}_3 = [-3, -1]^T$, $\boldsymbol{x}_4 = [-2, 0]^T$, $\boldsymbol{x}_5 = [2, 0]^T$, $\boldsymbol{x}_6 = [3, 1]^T$, $\boldsymbol{x}_7 = [4, 0]^T$, $\boldsymbol{x}_8 = [3, -1]^T$, $\boldsymbol{x}_9 = [-1, 7]^T$, $\boldsymbol{x}_{10} = [0, 8]^T$, $\boldsymbol{x}_{11} = [1, 7]^T$, $\boldsymbol{x}_{12} = [0,\ 6]^T$.

a. Let $m = 2$. Consider the vectors $\boldsymbol{w}_1 = [0, 0]^T$ and $\boldsymbol{w}_2 = [0, 7]^T$, such that the first one represents the points $\boldsymbol{x}_1$ through $\boldsymbol{x}_8$ and the second one represents the rest of the points in X. Compute the values of Γ and $\hat{\Gamma}$ (Section 16.4.1).

b. Let $m = 3$. Consider the vectors $\boldsymbol{w}_1 = [-3, 0]^T$ $\boldsymbol{w}_2 = [3, 0]^T$, and $\boldsymbol{w}_3 = [0, 7]^T$, so that the first one represents the points $\boldsymbol{x}_1$ through $\boldsymbol{x}_4$, the second represents the points $\boldsymbol{x}_5$ through $\boldsymbol{x}_8$, and the third represents the rest of the points of X. Compute the values of Γ and $\hat{\Gamma}$.

c. Let $m = 4$. Define $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ as in the previous case. Also, define $\boldsymbol{w}_3 = [-0.5, 7.5]^T$ and $\boldsymbol{w}_4 = [0.5, 6.5]^T$, so that the first represents $\boldsymbol{x}_9$ and $\boldsymbol{x}_{10}$, while the second represents $\boldsymbol{x}_{11}$ and $\boldsymbol{x}_{12}$. Compute the values of Γ and $\hat{\Gamma}$.

d. What conclusions can you draw from the comparison of the values of Γ and $\hat{\Gamma}$ obtained from the preceding three cases?

16.7 Estimate the number of operations required for the computation of Dunn's index, D_m, given by Eq. (16.19). What is the total number of computations required for the computation of D_m, for $m = 1, \ldots, N$?

16.8 Define explicitly $diam_i^{\mathrm{GG}}$ and $diam_i^{\mathrm{RNG}}$ that are involved in the definitions of the GG and the RNG Dunn-like indices. Using these definitions derive explicitly the GG and the RNG Dunn-like indices.

16.9 Show that $D_m^{\mathrm{GG}} \leq D_m^{\mathrm{RNG}} \leq D_m^{\mathrm{MST}}$.

Hint: Use the fact that for a cluster C_i, $E_i^{\mathrm{MST}} \subseteq E_i^{\mathrm{RNG}} \subseteq E_i^{\mathrm{GG}}$.

16.10 **a.** Show that the R_{ij}^{MST} given by Eq. (16.26) satisfies the conditions (C1)–(C5).

b. Taking into account the definition of R_{ij}^{MST}, define R_{ij}^{GG} and R_{ij}^{RNG} and show that they also satisfy the conditions (C1)–(C5).

16.11 Show that $DB_m^{\mathrm{GG}} \geq DB_m^{\mathrm{RNG}} \geq DB_m^{\mathrm{MST}}$.

Hint: Use the fact that for a cluster C_i, $E_i^{\mathrm{MST}} \subseteq E_i^{\mathrm{RNG}} \subseteq E_i^{\mathrm{GG}}$.

16.12 Explain intuitively why the MST DB is more robust to the presence of noisy vectors than the original DB.

16.13 **a.** Prove that, as $q \rightarrow 1^+$, PC and PE tend to 1 and 0, respectively.

b. Prove that, as $q \rightarrow \infty$, PC and PE tend to $1/m$ and $\log_a m$, respectively.

c. Show that in the latter case, in the plots of PC and PE, the most significant knee is exhibited at $m = 2$.

16.14 Prove that, as $q \rightarrow \infty$, the XB index tends to ∞, while XB_q becomes indeterminate.

Hint: Use the following facts: (a) $\lim_{q\to\infty} \boldsymbol{w}_i = \boldsymbol{w}$, where $\boldsymbol{w}$ is the mean vector of all vectors in X, $i = 1, \ldots, m$, and (b) for $q \to \infty$, $u_{ij} = 1/m$.

16.15 **a.** Prove that $\lim_{q\to 1^+} FS_q = 2N\text{trace}(S_w) - N\text{trace}(S_m)$, where S_w and S_m are the within and the mixture scatter matrices defined in Chapter 5.

b. Prove that $\lim_{q\to\infty} FS_q = 0$.

Hint: Use the hints given in the previous problem.

16.16 Prove that the distance of a point $\boldsymbol{x}_i$ from a sphere with center $\boldsymbol{c}_j$ and radius r_j is given by Eq. (16.48).

16.17 Consider a cluster C that consists of the points of a circular arc of radius r, subtending an angle ϕ (of course, this case is of theoretical interest, since the number of vectors in C would be infinite). The covariance matrix of this arc is

$$\Sigma = \frac{1}{L_\phi} \int_{-\phi/2}^{\phi/2} \boldsymbol{x}\boldsymbol{x}^T dl - \boldsymbol{m}\boldsymbol{m}^T \tag{16.65}$$

where $\boldsymbol{x} = [r\cos\theta, r\sin\theta]^T$ is a point on the arc, $dl = rd\theta$, and L_ϕ is the arc length. (a) Prove that

$$\delta = \frac{\phi}{2\pi\sqrt{1 - \frac{4sin^2(\phi/2)}{\phi^2}}} \tag{16.66}$$

What is the value of δ when $\phi = 2\pi$?

(b) Repeat for the case where the length of the cluster is a line segment of length L.

16.18 Consider a square of side a. Consider a grid of horizontal and vertical lines in the square so that the distance between two adjacent parallel lines is r. Place in it $(a/2r)^2$ vectors (of course, $a/2r$ is assumed to be an integer) such that each of them lies at an intersection point of a grid and the circles of radius r centered at these points do not intersect at more than one point. (a) Compute the packing density of the square. Repeat the above for the case where r is replaced by $r/2$.

(b) Assuming that no circle is allowed to have a part of it outside the square, is it possible to determine an arrangement of points in the square that results in a higher packing density?

16.19 Sometimes we say that the Hopkins test includes "first-order" information on the data set X and the Cox–Lewis test "second-order" information. Can you justify this proposition?

16.20 Repeat Example 16.10 and explain why (a) the number of decomposition layers is greater in clustered data than in random data and (b) the number of decomposition layers is greater in random data than in regular data.

REFERENCES

[Akai 85] Akaike H., "Prediction and Entropy," in *A Celebration of Statistics* (Atkinson A.C., Fienberg S.E., eds.), Sprieger-Verlag, New York, pp. 1–24, 1985.

[Back 81] Backer E., Jain A.K. "A clustering performance measure based on fuzzy set decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 3(1), pp. 66–75, 1981.

[Bail 82] Bailey T.A., Dubes R.C. "Cluster validity profiles," *Pattern Recognition*, Vol. 15, pp. 61–83, 1982.

[Bake 74] Baker F.B. "Stability of two hierarchical grouping techniques—Case 1. Sensitivity to data errors," *Journal of the American Statistical Association*, Vol. 69, 440–445, 1974.

[Bake 76] Baker F.B., Hubert L.J. "A graph-theoretic approach to goodness of fitting complete-link hierarchical clustering," *Journal of the American Statistical Association*, Vol. 71, pp. 870–878, 1976.

[Bart 62] Barton D.E., David F.N. "Randomization basis for multivariate tests in the bivariate case—randomness of points in the plane," *Bulletin of the International Statistical Institute*, Vol. 39, pp. 455–467, 1962.

[Beni 94] Beni G., Liu X. "A least biased fuzzy clustering method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16(9), pp. 954–960, 1994.

[Besa 73] Besag J.E., Gleaves J.T. "On the detection of spatial pattern in plant communities," *Bulletin of International Statistics Institute*, Vol. 45, p. 153, 1973.

[Bezd 74] Bezdek J.C. "Cluster validity with fuzzy sets," *Journal of Cybernetics*, Vol. 3(3), pp. 58–72, 1974.

[Bezd 75] Bezdek J.C. "Mathematical models for systematics and taxonomy," in *Proc. 8th Int. Conf. in Numerical Taxonomy* (Estarook G., ed.), Freeman, San Francisco pp. 143–166, 1975.

[Boug 04] Bouguessa M., Wang S-R. "A new efficient validity index for fuzzy clustering," *Proceedings of the 3rd International Conference on Machine Learning and Cybernetics*, pp. 1914–1919, 2004.

[Bout 04] Boutin F., Hascoët M. "Cluster validity indices for graph partitioning," *Proceedings of the 8th Conference on Information Visualization*, pp. 376–381, 2004.

[Cono 79] Conover W.J., Benent T.R., Iman R.L. "On a method for detecting clusters of possible uranium deposits," *Technometrics*, Vol. 21, pp. 277–282, 1979.

[Cox 76] Cox T.F., Lewis T. "A conditioned distance ratio method for analyzing spatial patterns," *Biometrika*, Vol. 63, p. 483, 1976.

[Cunn 72] Cunningham K.M., Ogilvie J.C. "Evaluation of hierarchical grouping techniques: A preliminary study," *Computer Journal*, Vol. 15, pp. 209–213, 1972.

[Dave 90] Dave R.N., Patel K.J. "Progressive fuzzy clustering algorithms for characteristic shape recognition," *Proc. North American Fuzzy Inf. Process. Soc. Workshop*, Toronto, pp. 121–124, 1990.

[Davi 79] Davies D.L., Bouldin D.W. "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 1(2), pp. 224–227, 1979.

[Diac 83] Diaconis P., Efron B. "Computer-intensive methods in statistics," *Scientific American*, May, pp. 116–130, 1983.

[Digg 83] Diggle P.J. *Statistical Analysis of Spatial Point Patterns*, Academic Press, 1983.

[Dube 79] Dubes R.C., Jain A.K. "Validity studies in clustering methodologies," *Pattern Recognition*, Vol. 11, pp. 235–254, 1979.

[Dube 87a] Dubes R.C. "How many clusters are best? An experiment," *Pattern Recognition*, Vol. 20(6), pp. 645–663, 1987.

[Dube 87b] Dubes R.C., Zeng G. "A test for spatial homogeneity in cluster analysis," *Journal of Classification*, Vol. 4, pp. 33–56, 1987.

[Dunn 74] Dunn J.C. "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, Vol. 4, pp. 95–104, 1974.

[Dunn 76] Dunn J.C. "Indices of partition fuzziness and the detection of clusters in large data sets," *in Fuzzy Automata and Decision Processes* (Gupta M.M., ed.), Elsevier, 1976.

[Efro 79] Efron B. "Bootstrap methods: Another look at jackknife," *Applied Statistics*, Vol. 7, pp. 1–26, 1979.

[Farr 69] Farris J.S. "On the cophenetic correlation coefficient," *Systematic Zoology*, Vol. 18, pp. 279–285, 1969.

[Fill 71] Fillenbaum S., Rapoport A. *Structures in the Subjective Lexicon*, Academic Press, 1971.

[Fowl 83] Fowlkes E.B., Mallows C.L. "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, Vol. 78, pp. 553–569, 1983.

[Fral 98] Fraley C., Raftery A.E., "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, Vol. 41, No. 8, pp. 578–588, 1998.

[Frie 79] Friedman J.H., Rafsky L.C. "Multivariate generalization of the Wald–Wolfowitz and Smirnov two-sample tests," *Annual Statistics*, Vol. 7, pp. 697–717, 1979.

[Gath 89] Gath I., Geva A.B. "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11(7), pp. 773–781, 1989.

[Grei 64] Greig-Smith P. *Quantitative Plant Ecology*, 2nd ed., Butterworth, 1964.

[Gord 99] Gordon A. *Classification, 2nd edition*, Chapman and Hall/CRC press, London, 1999.

[Halk 00] Halkidi M., Vazirgiannis M., Batistakis Y. "Quality scheme assessment in the clustering process," *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 265–276, 2000.

[Halk 01] Halkidi M., Vazirgiannis M. "Clustering validity assessment: finding the optimal partitioning of a data set," *Proceedings of the International Conference of Data Mining 2001*, pp. 187–194, 2001.

[Halk 02a] Halkidi M., Batistakis Y., Vazirgiannis M. "Cluster validity methods: part 1," *SIGMOD Record*, Vol. 31(2), pp. 40–45, 2002.

[Halk 02b] Halkidi M., Batistakis Y., Vazirgiannis M. "Cluster validity methods: part 2," *SIGMOD Record*, Vol. 31(3), pp. 19–27, 2002.

[Hart 75] Hartigan J.A. *Clustering Algorithms*, John Wiley & Sons, 1975.

[Hast 01] Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning*, Springer, 2001.

[Hoff 87] Hoffman R.L., Jain A.K. "Sparse decompositions for exploratory pattern analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9(4), pp. 551–560, 1987.

[Hopk 54] Hopkins B. "A new method for determining the type of distribution of plant-individuals," *Annals of Botany*, Vol. 18, pp. 213–226, 1954.

[Hube 74] Hubert L.J. "Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures," *Journal of the American Statistical Association*, Vol. 69, pp. 698–704, 1974.

[Hube 76] Hubert L.J., Schultz J. "Quadratic assignment as a general data-analysis strategy," *British Journal of Mathemetical and Statistical Psychology*, Vol. 29, pp. 190–241, 1976.

[Hube 85] Hubert L.J., Arabie P. "Comparing partitions," *Journal of Classification*, Vol. 2, pp. 193–218, 1985.

[Hurv 89] Hurvich C.M., Tsai C-L, "Regression and time series model selection in small samples," *Biometrika* Vol. 76, pp. 297–307, 1989.

[Ivch 1991] Ivchenko G., Medvedev Y., Chistyakov A. *Problems in Mathematical Statistics*, Mir Publishers, Moscow, 1991.

[Jain 87a] Jain A.K., Dubes R., Chen C.C. "Bootstrapping techniques for error estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 9, pp. 628–633, 1987.

[Jain 87b] Jain A.K., Moreau J.V. "Bootstrap technique in cluster analysis," *Pattern Recognition*, Vol. 20(5), pp. 547–568, 1987.

[Jain 88] Jain A.K., Dubes R.C. *Algorithms for Clustering Data*, Prentice Hall, 1988.

[Kauf 90] Kaufman L., Rousseeuw P. *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley New York, 1990.

[Kirl 00] Kirlin R.L., Dizaji R.M., "Cluster order using clustering performance index rate, CPIR," NORSIG 2000, Kolmarden, Sweden, June 2000.

[Kris 93] Krishnapuram R., Frigui H., Nasraoui O. "Quadratic shell clustering algorithms and the detection of second-degree curves," *Pattern Recognition Letters*, Vol. 14(7), pp. 545–552, July 1993.

[Kris 95a] Krishnapuram R., Frigui H., Nasraoui O. "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation—Part I," *IEEE Transactions on Fuzzy Systems*, Vol. 3(1), pp. 29–43, 1995.

[Kris 95b] Krishnapuram R., Frigui H., Nasraoui O. "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation—Part II," *IEEE Transactions on Fuzzy Systems*, Vol. 3(1), pp. 44–60, 1995.

[Lang 98] Langan D.A., Modestino J.W., Zhang J. "Cluster validation for unsupervised stochastic model-based image segmentation," *IEEE Transactions on Image Processing*, Vol. 7(2), pp. 180–195, 1998.

[Ling 72] Ling R.F. "On the theory and construction of *k*-clusters," *Computer Journal*, Vol. 15, pp. 326–332, 1972.

[Ling 73] Ling R.F. "Probability theory of cluster analysis," *Journal of the American Statistical Association*, Vol. 68, pp. 159–164, 1973.

[Ling 76] Ling R.F., Killough G.S. "Probability tables for cluster analysis based on a theory of random graphs," *Journal of the American Statistical Association*, Vol. 71, pp. 293–300, 1976.

[Lu 00] Lu X. "Comparisons among information-based criteria, a novel modification thereof, and the Monte Carlo Markov chain method," MSc Thesis, University of Victoria, British Columbia, Canada, July 2000.

[Mant 67] Mantel N. "The detection of disease clustering and a generalized regression approach," *Cancer Research*, Vol. 27, pp. 209–220, 1967.

[Mead 74] Mead R. "A test for spatial pattern at several scales using data from a grid of contiguous quadrats," *Biometrics*, Vol. 30, pp. 295–308, 1974.

[Mill 80] Milligan G.W. "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrica*, Vol. 45, pp. 325–342, 1980.

[Mill 83] Milligan G.W., Soon S.C., Sokol L.M. "The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, pp. 40–47, 1983.

[Mill 85] Milligan G.W., Cooper M.C. "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, Vol. 50, pp. 159–179, 1985.

[Naus 82] Naus J.J. "Approximations for distributions of scan statistics," *Journal of the American Statistical Association*, Vol. 77, pp. 177–183, 1982.

[Neym 72] Neyman J., Scott E.L. "Processes of clustering and applications," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications* (Lewis P.A.W., ed.), John Wiley & Sons, 1972.

[Pal 95] Pal N.R., Bezdek J.C. "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, Vol. 3(3), pp. 370–379, 1995.

[Pal 97] Pal N.R., Biswas J. "Cluster validation using graph theoretic concepts," *Pattern Recognition*, Vol. 30(6), pp. 847–857, 1997.

[Pana 83] Panayirci E., Dubes R.C. "A test for multidimensional clustering tendency," *Pattern Recognition*, Vol. 16(4), pp. 433–444, 1983.

[Papo 91] Papoulis A. *Probability, Random Variables and Stochastic Processes*, 3rd ed., McGraw-Hill, 1991.

[Piel 69] Pielou E.C. *An Introduction to Mathematical Ecology*, John Wiley & Sons, 1969.

[Post 93] Postaire J.G., Zhang R.D., Lecocq-Botte C. "Cluster analysis by binary morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15(2), pp. 170–180, 1993.

[Ripl 77] Ripley B.D. "Modelling spatial patterns," *Journal of the Royal Statistical Society*, Vol. B39, pp. 172–212, 1977.

[Ripl 78] Ripley B.D., Silverman B.W. "Quick tests for spatial interaction," *Biometrika*, Vol. 65, pp. 641–642, 1978.

[Ripl 81] Ripley B.D. *Spatial Statistics*, John Wiley & Sons, 1981.

[Riss 78] Rissanen J. "Modeling by shortest data description," *Automatica* 14, pp. 465–471, 1978.

[Riss 89] Rissanen J. "Stochastic complexity in statistical enquiry," *Series in computer science*, 15, World Scientific, Singapore, 1989.

[Rolp 68] Rolph F.J., Fisher D.R. "Tests for hierarchical structure in random data sets," *Systematic Zoology*, Vol. 17, pp. 407–412, 1968.

[Rolp 70] Rolph F.J. "Adaptive hierarchical clustering schemes," *Systematic Zoology*, Vol. 19, pp. 58–82, 1970.

[Schw 76] Schwarz G. "Estimating the dimension of a model," *Annals of Statistics* Vol. 6, pp. 461–464, 1976.

[Sclo 87] Sclove S.L. "Application of model-selection criteria to some problems in multivariate analysis," *Psychometrika*, Vol. 52, pp. 333–343, 1987.

[Sent 07] Sentelle C., Hong S.L., Georgiopoulos M., Anagnostopoulos G.C. "A fuzzy gap statistic for fuzzy c-means," *Proceedings of the 11th IASTED International Conference on Artificial Intelligence and Soft Computing*, pp. 68–73, 2007.

[Shar 96] Sharma S. *Applied Multivariate Techniques*, John Wiley & Sons Inc., 1996.

[Shre 64] Shreider Y.A. *Method of Statistical Testing: Monte Carlo Method*, Elsevier North-Holland, 1964.

[Silv 78] Silverman B., Brown T. "Short distances, flat triangles and Poisson limits," *Journal of Applied Probability*, Vol. 15, pp. 815–825, 1978.

[Smit 84] Smith S.P., Jain A.K. "Testing for uniformity in multidimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, pp. 73–81, 1984.

[Snea 77] Sneath P.H.A. "A significance test for clusters in UPGMA phenograms obtained from squared Euclidean distance," *Classification Soc. Bulletin*, Vol. 4, pp. 2–14, 1977.

[Sobo 84] Sobol I.M. *The Monte Carlo Method*, Mir Publishers, Moscow, 1984.

[Stra 75] Strauss D.J. "Model for clustering," *Biometrika*, Vol. 62, pp. 467–475, 1975.

[Tibs 01] Tibshirani R., Walther G., Hastie T. "Estimating the number of clusters in a data set via the gap statistic," *Journal of Royal Statistics Society B*, Vol. 63, pp. 411–423, 2001.

[Wind 81] Windham A.P. "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets and Systems*, Vol. 5, pp. 177–185, 1981.

[Wind 82] Windham M.P. "Cluster validity for the fuzzy c-means clustering algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 4(4), pp. 357–363, July 1982.

[Xie 91] Xie X.L., Beni G. "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13(8), pp. 841–846, 1991.

[Yarm 87] Yarman-Vural F., Ataman E. "Noise, histogram and cluster validity for Gaussian mixtured data," *Pattern Recognition*, Vol. 20(4), pp. 385–401, 1987.

[Zeng 85] Zeng G., Dubes R.C. "A comparison of tests for randomness," *Pattern Recognition*, Vol. 18(2), pp. 191–198, 1985.