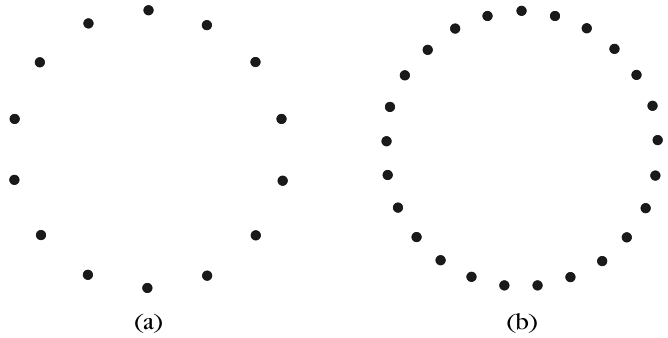

FIGURE 16.8

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.

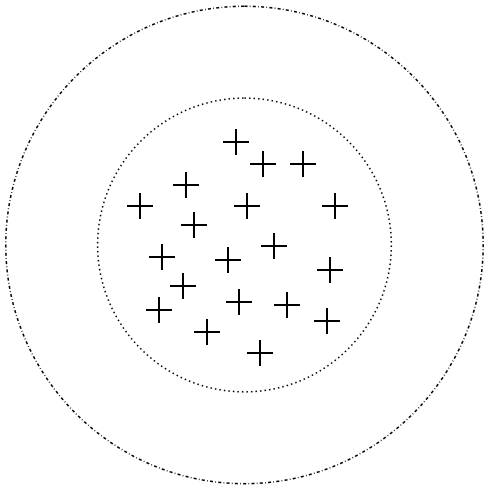

FIGURE 16.9

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

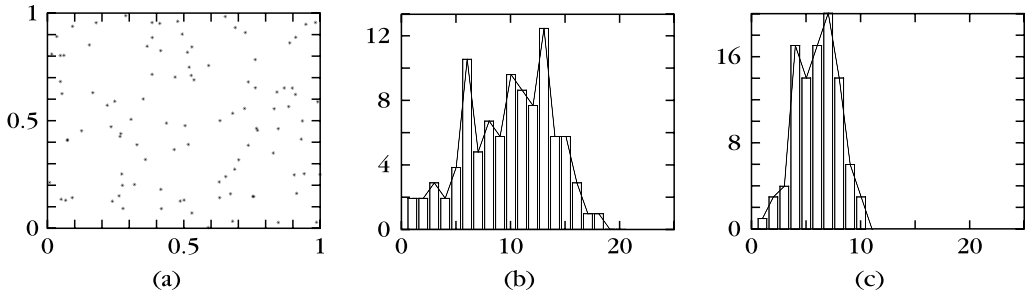

FIGURE 16.10

(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

$$\rho = \frac{L}{V} V_r \tag{16.55}$$

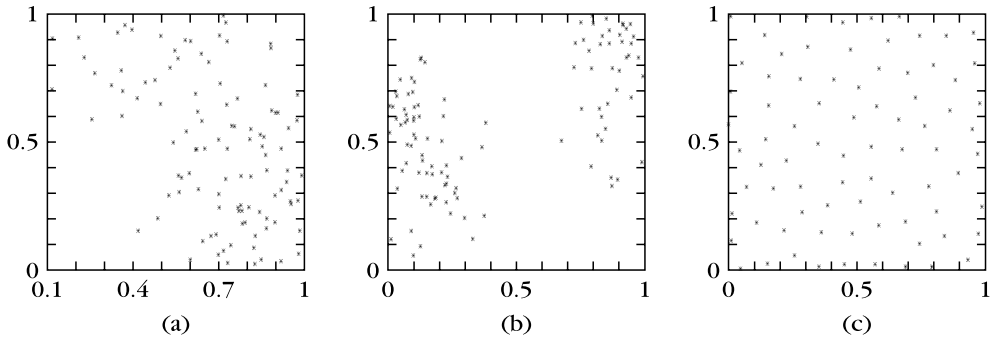

FIGURE 16.11

(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

$$\text{var}(q|e, H_0) = \frac{2MN}{L(L-1)}\left[\frac{2MN-L}{L} + \frac{e-L+2}{(L-2)(L-3)}[L(L-1)-4MN+2]\right] \tag{16.59}$$

where $L = M + N$. Moreover, it can be shown [Frie 79] that if $M, N \to \infty$ 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:

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

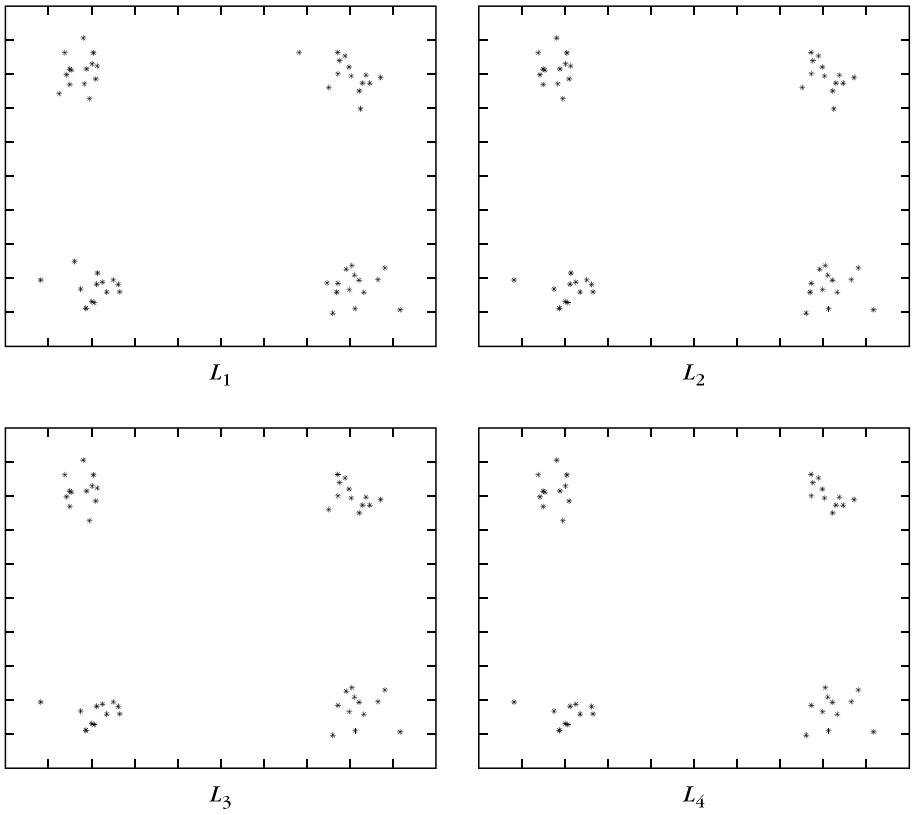

FIGURE 16.12

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

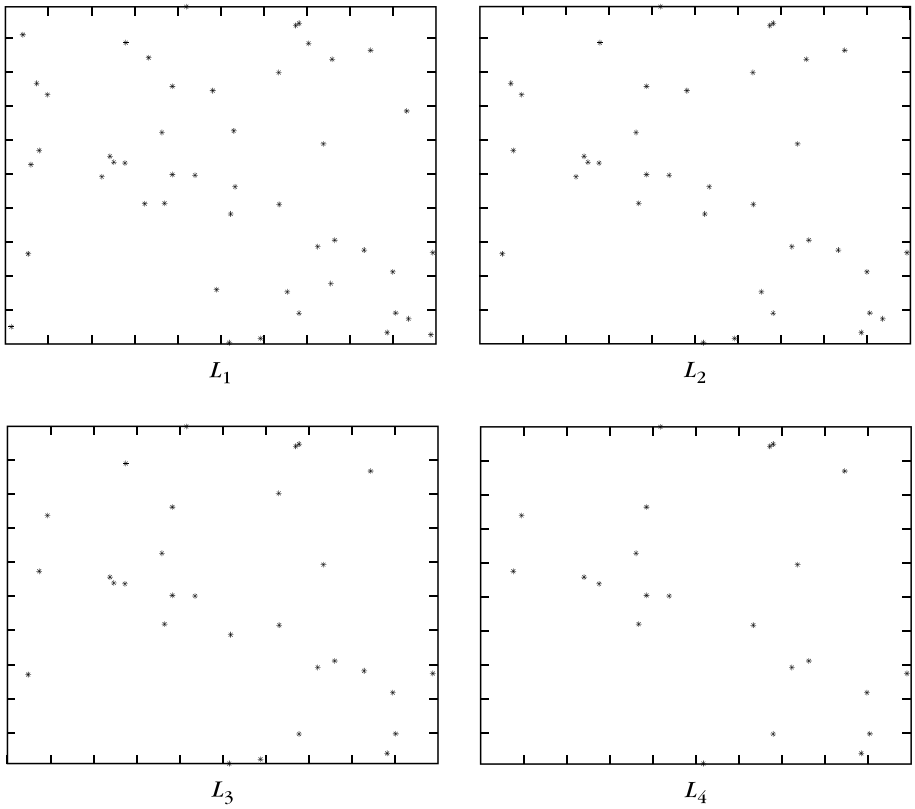

FIGURE 16.13

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 $0.5^2 I$.